NASA Contractor Report 195471

Information Switching Processor (ISP) Contention Analysis and Control

Thomas Inukai
COMSAT Laboratories
Clarksburg, Maryland

July 1995

Prepared for
Lewis Research Center
Under Contract NAS3-25933

National Aeronautics and
Space Administration

TECHNICAL SUPPORT FOR DIGITAL SYSTEM TECHNOLOGY DEVELOPMENT

Task Order No.1

Final Report

INFORMATION SWITCHING PROCESSOR (ISP) CONTENTION ANALYSIS AND CONTROL

Submitted to

National Aeronautics and Space Administration
Lewis Research Center
2100 Brookpark Road
Cleveland, Ohio 44135

Contract No. NAS3-25933

February 10, 1992

COMSAT LABORATORIES
22300 COMSAT DRIVE, CLARKSBURG, MARYLAND 20871

Table of Contents

Table of Contents (cont'd)

List of Illustrations

List of Illustrations (cont'd)

List of Tables

1 Introduction

Future satellite communications, as a viable means of communications and an alternative to terrestrial networks, demand flexibility and low end-user cost. On-board switching/processing satellites potentially provide these features, allowing flexible interconnection among multiple spot beams, direct to the user communications services using very small aperture terminals (VSATs), independent uplink and downlink access/transmission system designs optimized to user's traffic requirements, efficient TDM downlink transmission, and better link performance. A flexible switching system on the satellite in conjunction with low-cost user terminals will likely benefit future satellite network users.

In designing a satellite system with on-board processing, the selection of a switching architecture is often critical. The on-board switching function can be implemented by circuit switching or destination-directed packet switching, which is also known as fast packet switching. 'Destination-directed packet switching has several attractive features, such as self-routing without on-board switch reconfiguration, no switch control memory requirements, efficient bandwidth utilization for packet switched traffic, and accommodation of circuit switched traffic. These advantages have been fully described in various papers in the past [1] - [4].

Destination-directed packet switching, however, has two potential concerns: (a) contention and (b) congestion. Contention occurs when two or more packets from different input ports attempt to reach the same output port at the same time, and congestion occurs when an on-board buffer overflows due to the limitation in switch routing capability or downlink transmission capacity. This report specifically deals with the first problem. It includes a description and analysis of various self-routing switch structures, the nature of contention problems, and contention resolution techniques. The following is a brief description of the contents of this report.

Section 2 describes the satellite network requirements which are the basis of this study and includes a reference network architecture, on-board baseband processor configuration, and problem statement.

Section 3 presents contention-free switch architectures and contention-based architectures. Contention-based architectures include three types of contention resolution techniques, such as output port reservation, path setups prior to packet routing, and address filtering. Simulation results on switch throughput performance are also provided.

Section 4 addresses multiplexing schemes at the switch output ports for circuit and packet switched traffic. The use of a shared buffer for the two types of traffic potentially reduces on-board packet congestion.

Section 5 considers the feasibility of integrating circuit and packet switching with a fast packet switch. This type of switch is more flexible than separate switches for circuit and packet switched traffic.

Section 6 briefly describes possible techniques for congestion control. Since this subject is not a part of this study and also requires extensive investigation, no detailed analysis is provided.

Section 7 summarizes the study results and presents recommendations for future study.

2 Satellite Network Requirements

This section describes the reference network architecture and on-board processor (OBP) configuration used for this study. Although the system architecture assumes specific network parameters, the general discussions, results, and conclusions presented in the following sections are not restricted to the particular sample architecture and are applicable to a destination-directed packet switched system in general. A description of the study task is also included in this section.

2.1 Reference Network Architecture

The satellite network under consideration operates at the 30/20-GHz frequency band and provides flexible, low-cost mesh VSAT services to users located in the continental United States (CONUS). The satellite antenna coverage consists of eight fixed uplink beams, eight hopping downlink beams, and an intersatellite link (ISL), where each downlink beam has eight dwell locations. An on-board baseband processor (OBP) provides connectivity among uplink and downlink beams. Figure 1 depicts the system concept.

The system provides voice, data, facsimile, datagram, teleconferencing, and video communications services. To support these services, the system incorporates two types of transmission modes. The first type is a continuous transmission of circuit switched traffic at 2.048 Mbit/s, which is trunked to either single or multiple destination stations and does not require on-board demultiplexing of individual channels. The second type is also continuous transmission at 64 kbit/s, but it consists of fixed length packets with variable destination stations. The satellite routes these packets to the proper downlink beams according to the routing information contained in the packet headers. Each uplink beam supports forty (40) 2.048-Mbit/s and one thousand and twenty-four (1,024) 64-kbit/s FDMA carriers. In addition, each uplink beam includes one or two 64-kbit/s time-slotted signaling channels operating either in TDMA or random access mode. These channels are shared by all user earth stations within a beam to send orderwire messages to the satellite, such as capacity allocation/deallocation requests, traffic types, traffic characteristics, station status, and other messages necessary for network operation.

Downlink transmission to each beam is burst TDM at 150 Mbit/s and consists of eight TDM bursts, each destined to one of eight dwell locations within the beam. The circuit switched traffic and packet switched traffic are multiplexed on board the satellite to form a single TDM burst per dwell area for efficient satellite power utilization and simpler user earth station processing. Downlink orderwire messages for capacity assignment and station control are also included in the burst. For broadcast and multicast operation, the satellite must be capable of duplicating and transmitting the received message to up to 64 downlink dwell locations.

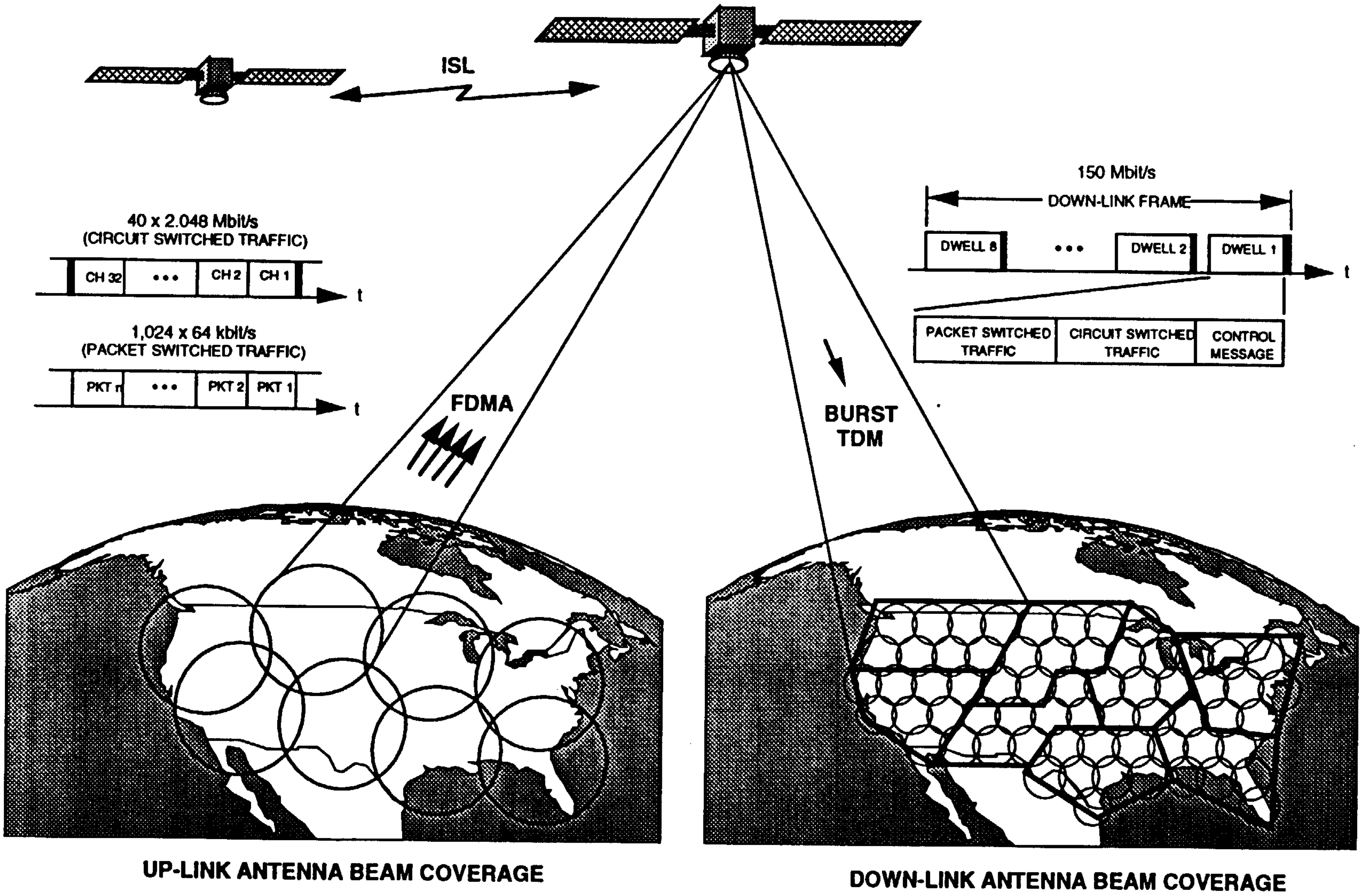


Figure 1. Reference Network Architecture

For the purpose of this study, it is assumed that the ISL transmission capacity is the same as that of one beam, i.e., about 150 Mbit/s, and includes both circuit and packet switched traffic.

The total system capacity of 2.048-Mbit/s trunk service is 737 Mbit/s (= 2.048 Mbit/s/carrier x 40 carriers/beam x 9 beams including ISL), and that of 64-kbit/s packet service is 590 Mbit/s (= 64 kbit/s/carrier x 1,024 carriers/beam x 9 beams including ISL).

2.2 On-Board Baseband Processor Configuration

Connectivity among spot beams is established by the OBP of which a functional block diagram is shown in Figure 2. The 2.048 Mbit/s uplink carriers from each beam are demultiplexed and demodulated by a multicarrier demultiplexer/demodulator (MCDD), FEC decoded, frame synchronized, and suitably reformatted for subsequent switching. The circuit switch provides a routing path from an input port to one or more output ports, and the path configuration remains unchanged for the duration of a circuit switched call.

The packet switched traffic transmitted on a 64-kbit/s carrier is first demultiplexed and demodulated by an MCDD as in the previous case. The transmitted packets are detected by a packet synchronizer, FEC decoded, and assembled to form complete packets prior to routing. The (fast) packet switch routes these packets to the proper output ports according to the information contained in the packet headers.

The input processing functions can be implemented in several ways in an actual system. For example, TDM frame synchronization may be performed prior to FEC decoding to reduce FEC decoder complexity operating in a time-shared manner. This will, however, require a longer unique word to identify a TDM frame marker or separate FEC coding. The locations of packet synchronization and packet assembly functions can be interchanged for the same reason. This will also provide added protection on the packet headers with double FEC coding. The switching function can be performed by two independent switches as shown in the figure or can be implemented by a single integrated fast packet switch. This issue is further exploited in a later section.

The downlink processing functions include multiplexing of two types of traffic along with station control messages, burst TDM formatting, FEC encoding, and modulation. Typically, one burst per dwell location is transmitted to each beam in one frame period. The packet switched traffic is statistical, and the amount of traffic flow to a particular downlink beam changes from frame to frame, causing a potential buffer overflow. To minimize on-board packet congestion, a downlink buffer can be shared by circuit and packet switched traffic such that when the volume of circuit switched traffic is low, the excess buffer can be used for packet switched traffic on a contingency basis.

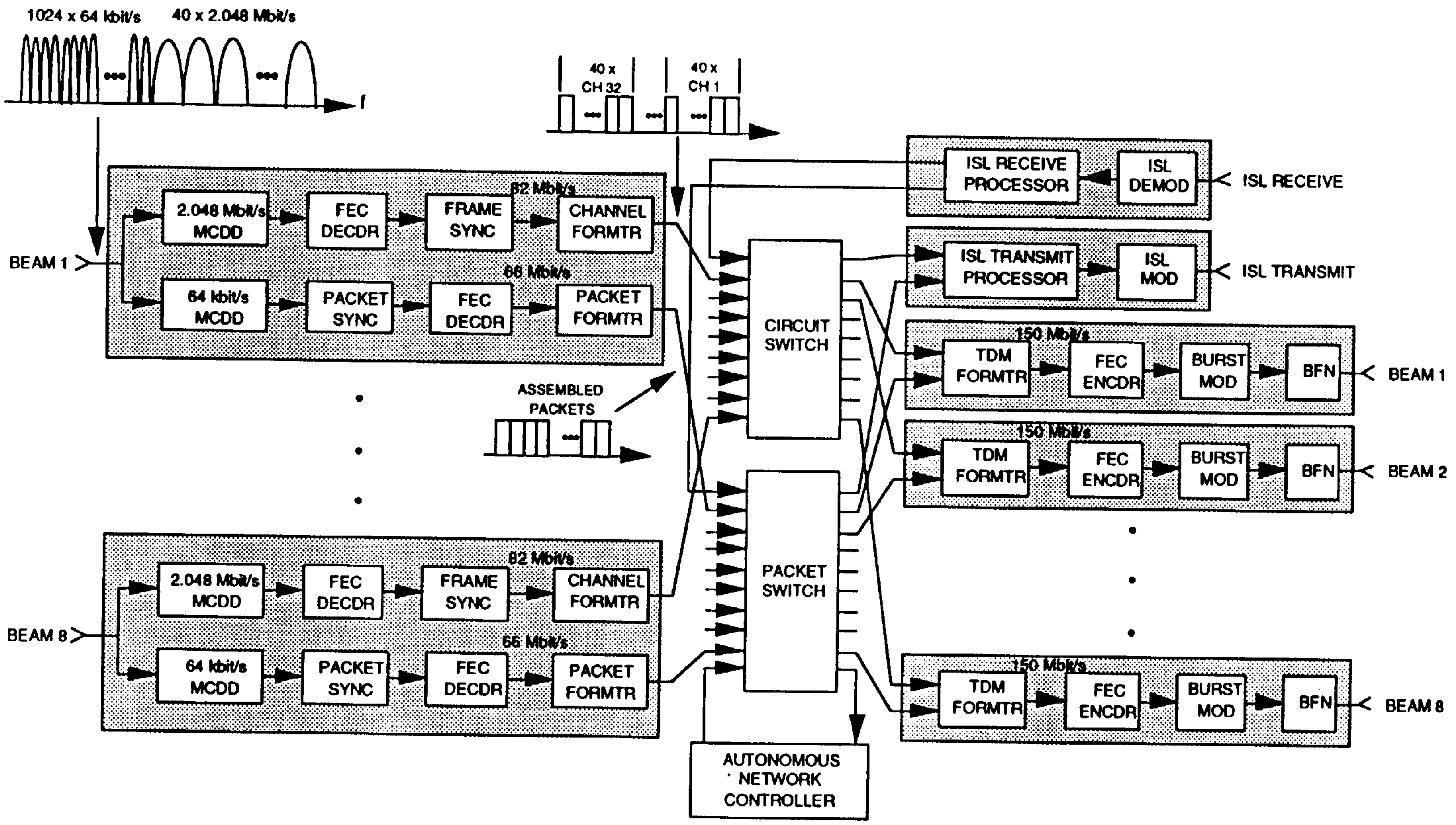


Figure 2. On-Board Baseband Processor Block Diagram

2.3 Study Task - Contention Control

There are two major system design issues associated with destination-directed packet switching. The first issue, which is the subject of this study, is a contention problem within a switch fabric. Since there will be no preassigned routing paths for data packets, a problem arises when packets from different input ports are to be routed to the same output port at the same time. This contention problem must be resolved by the use of a special switch structure or by some mechanism of avoiding simultaneous packet routing to the same output port.

The second issue is a congestion problem, which occurs when the total amount of packet switched traffic to some beam exceeds the allocated on-board buffer capacity. This is an inherent problem associated with virtually all fast packet switched systems, including broadband ISDN Asynchronous Transfer Mode (ATM) networks. Efficient flow/congestion control techniques must be devised to overcome this problem.

This study task deals with the contention problem. The following sections present several switch structures which are free from contention. In general, this type of switch architecture has a throughput limitation of a few gigabits per second. A higher capacity can be achieved with increased hardware complexity. Another type of switch architecture avoids contention by properly scheduling packet routing within a switching subsystem. A much higher throughput than the first type can be achieved with a moderate increase in control complexity. In this type of architecture, contention and congestion problems are inter-related, and contention-free switch operation is achieved at the expense of somewhat increased congestion. The report includes a detailed description of several such switching architectures, design tradeoffs, and a throughput analysis. Also included in the report are design approaches and contention/congestion control techniques for an integrated circuit/packet switch.

3 On-Board Switch Architectures and Contention

The difference between a circuit switch and a packet switch is that a packet switch performs like a statistical multiplexer while a circuit switch performs like a deterministic multiplexer. In a packet switch, several packets from different input ports may be destined to the same output port at the same time. This situation is referred as output contention (see Figure 3). Depending on the switch architecture, there are several means of resolving output contention.

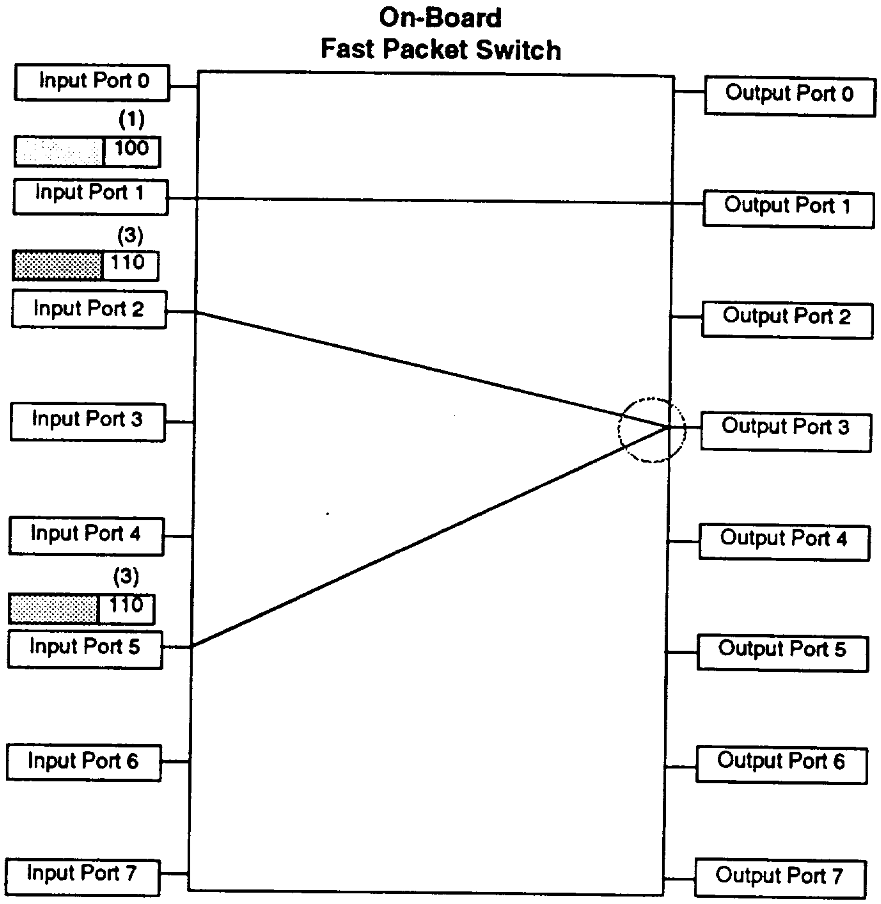


Figure 3. Output Contention

The switch architectures can be categorized into two classes: a contention-free switch and a contention switch. Three techniques of implementing a contention-free switch are described. Within the contention switch class, the switch architectures are classified according to the output contention resolution schemes. There are three subclasses: the first one employs an output reservation scheme at the input ports, the second one uses a path setup strategy to resolve blocking within a switching fabric and at the output ports at the same time, and the third one uses an address filter at the output port.

3.1 Contention-Free Switch Architectures

3.1.1 TDM Bus with Distributed Output Memories

The TDM bus is a degeneration of the banyan switch obtained by compressing the switching fabric into a bus (see Figure 4). In this scheme, all the packets from different input lines are multiplexed into a high-speed TDM bus. The speed of the TDM bus is the sum of the rates of the incoming lines. Since a TDM bus is a nonblocking switching fabric and the speed is N times faster than the link speed, the output port can receive up to N packets within one link slot time, where a link slot is defined as the ratio of the packet size and the link speed. Therefore, there is no output contention in the TDM bus with distributed output memories.

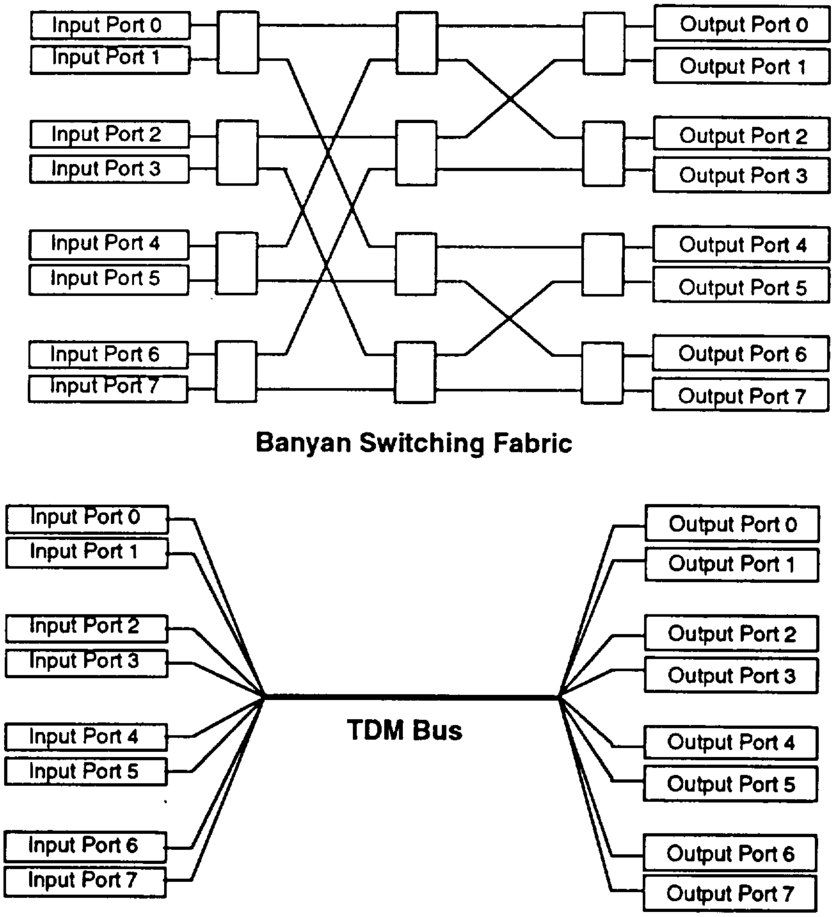


Figure 4. Correspondence Between Banyan Switch and TDM Bus Switch

One possible implementation of the TDM bus switch is described as follows. As shown in Figure 5, there are two separate logical buses within a physical bus; the first one is the packet (data payload) bus and the second the address (routing tag) bus. The address filter at each output port selects the desired packets on the TDM bus. Since

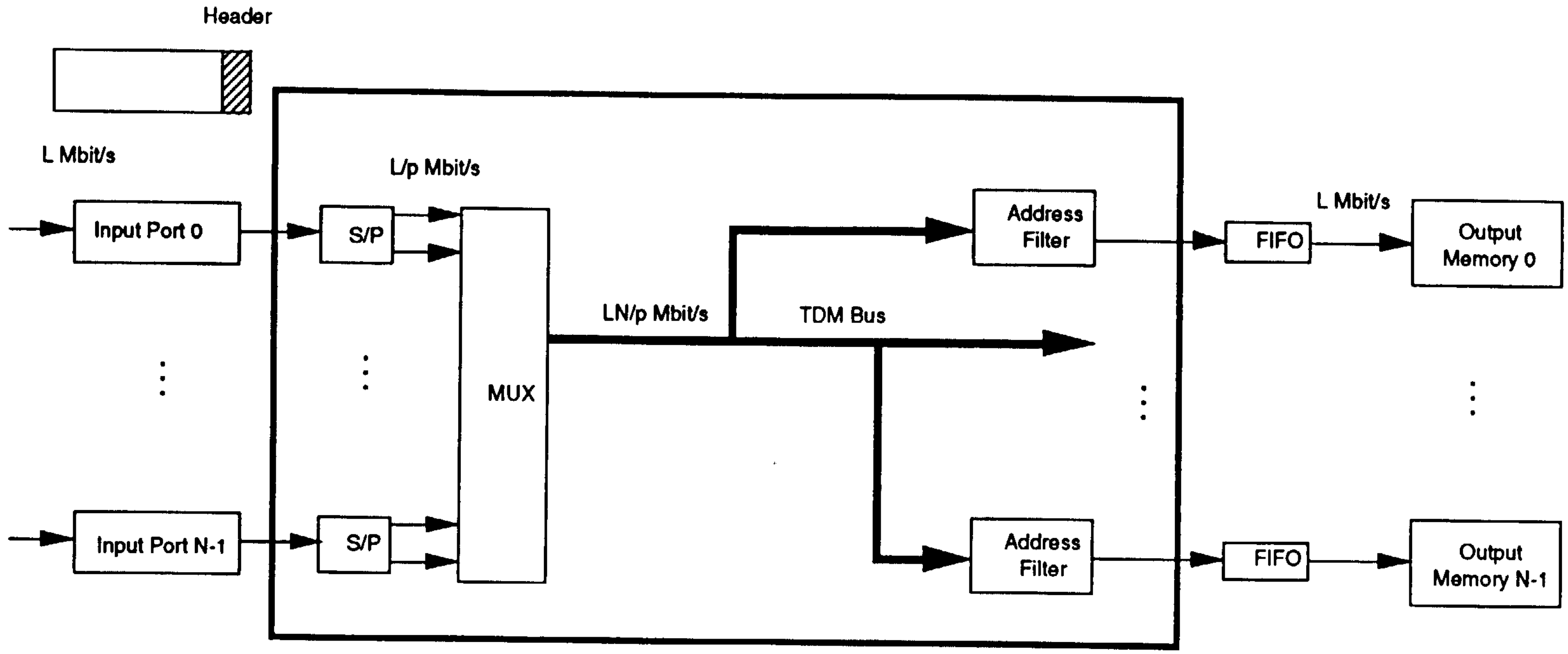


Figure 5. TDM Bus with Distributed Output Memories

more than one packet may arrive at the output port in one slot time, buffering is required at the output ports.

For point-to-point connection, the self-routing address (or the routing tag) requires at least Log_2 N bits. Since the TDM bus has an inherent broadcast capability, the TDM bus is also a point-to-multipoint nonblocking switching fabric. The multicast connection can be achieved by modifying the routing tag. The multicast routing tag requires N bits, where each bit represents one output port.

The packet filter structure depends on the addressing scheme. For a point-to-point addressing scheme, the packet filter is implemented using a comparator. For a multicast addressing scheme, the packet filter is a simple latch circuit.

The TDMA bus speed is given by LN/p, where L, N, and p are respectively the link speeed, the number of input ports, and the number of parallel bus lines. The general concerns with the TDM bus approach are the bus speed, the memory access time, and the bus loading (i.e., the number of input and output ports on the bus).

Congestion Issue

Since buffering is implemented at the output port, beam traffic congestion may occur at the output port. Congestion occurs when the incoming traffic is nonuniform in the output destination distribution or a short-term traffic intensity to a certain beam exceeds the beam capacity. The output buffer should be designed to absorb short-term fluctuations. When a buffer overflow occurs, some packets may be dropped.

3.1.2 Fiber Optic Ring Switch

The optic ring switch, as shown in Figure 6, uses the same design principle as the TDM bus, i.e., the optic bus speed is the sum of the rates of the incoming lines. The difference is that the optical ring can be operated in a much higher speed than the TDM electronic bus. Also, since the signal is regenerated at each port, the optic ring can accommodate more ports than the TDM bus.

The optic ring switch operates on a frame-by-frame basis. The autonomous network controller (ANC) periodically sends a frame marker to the bus. When an input port receives the frame marker, it inserts a packet with a routing tag into the preassigned empty slot. After the last input port has inserted a packet into an empty slot, the frame has been formed. The frame loaded with data packets is circulated around the output ports. There is an address filter attached to the bus at each output port. The filter is used to select the packets destined to the particular output port.

The optical ring switch has no internal blocking for point-to-point and point-to-multipoint connections and has no output contention.

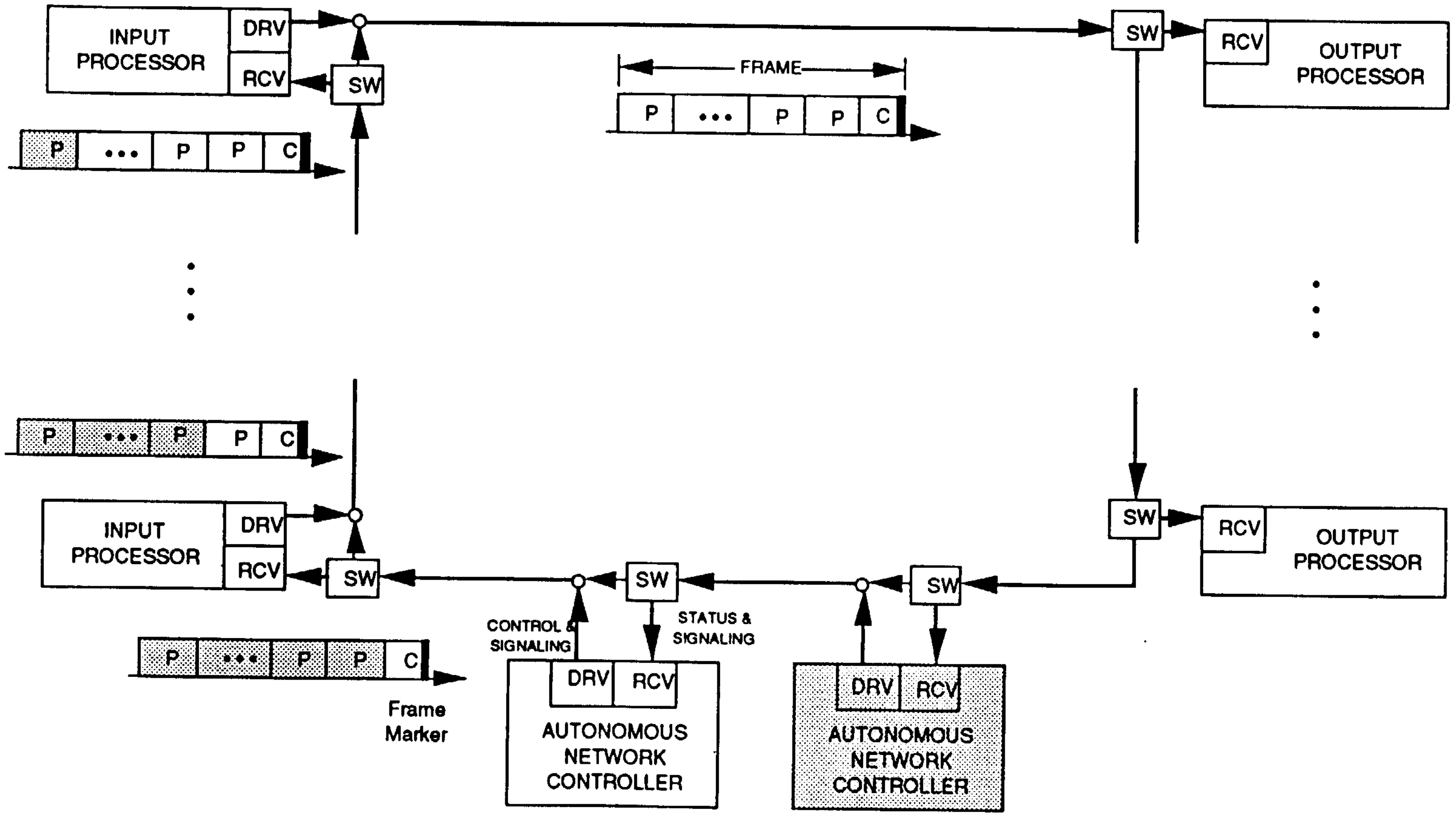


Figure 6. Fiber-Optic Ring Switch

Congestion Issue

Since buffering is also implemented at the output port of the switch as in the TDM bus, the same congestion problem as in the TDM bus exists.

3.1.3 Common Memory Switch

In this structure, all the packets from different input lines are multiplexed into a single TDM packet stream. The speed of the TDM stream is the sum of the incoming rates. The common memory approach, unlike the TDM bus switch with distributed output memories, shares one large memory among all the output ports.

There are several memory implementation techniques for switching. The simplest way, called a complete partition approach, is to partition the memory into N areas, where each area stores the packets destined to one output port. When packets arrive at the switch, the write controller examines the routing tag and stores the packet into the corresponding area sequentially. To provide contention-free operation, the size of the memory has to be at least N^2 packets, because each area needs to accommodate the worst situation that N packets are destined to the same output port at the same time. During the read cycle, the read controller reads packets sequentially from each area and sends the packets to the corresponding output ports through a demultiplexer. This approach is very similar to the TDM bus with distributed output memories. The disadvantage of this approach is that the memory is not shared efficiently.

Congestion Issue

As in the TDM bus, congestion occurs if the amount of traffic exceeds the capacity of the switch, i.e., the allocated area for each output port in the memory.

The second approach, called a complete sharing approach, is described below. The packets are stored in the common data memory, and the memory addresses of these packets are written into the control memory (see Figure 7). The self-routing addresses of the packets pass through the matched address filters and activate the corresponding pointer array. The address of the control memory is written into the pointer array according to the self-routing address. The packet control memory addresses whose packets go to the same output port are grouped into one array. The TDM output stream is formed by reading a packet out of the data memory for each output port using the address obtained from the control memory, while the address of the control memory is obtained using the addresses of each array corresponding to each output port. The packets on the TDM stream are demultiplexed into different output ports.

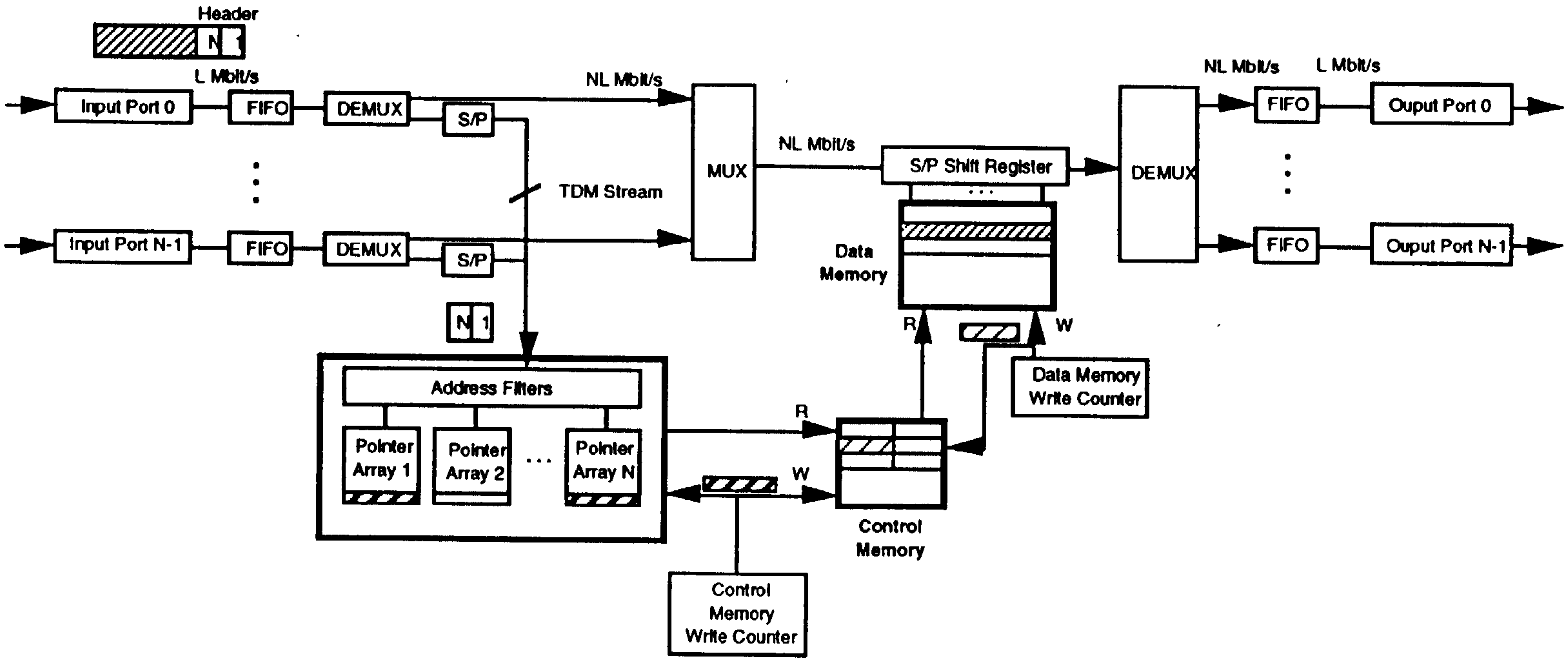


Figure 7. Common Memory Switch

Since the data memory and control memory are operated in a random read fashion, it is not easy to keep track of the empty memory space after the packets have been read out from the memory. Link list implementation of the memory is required to efficiently use the memory space. Each time a packet is read out from the memory, the address of the empty location enters an empty buffer pool. Each time a packet is written into the memory, an empty address is selected from the pool to store the packet.

For point-to-point connection, the shared-memory switch has no internal blocking and has no output contention.

The multicast operation is achieved using multiple writes to the pointer arrays since more than one pointer arrays will be activated at the same time for multicast connection. This is the most efficient multicast operation in terms of memory usage since there is only one copy of the multicast packet stored in the memory. Duplication of the packet is not performed on the packet itself instead on the memory address of the packet.

A concern with this approach is the memory access time requirement for high speed application. This problem can be overcome by using a wider parallel bus. Another concern is the memory size, which includes the data memory and the control memory.

Congestion Issue

Since memory is shared among all the output ports of the switch for the complete sharing scheme, the congestion problem is not as severe as the complete partitioning scheme. The memory acts as a very large buffer to absorb fluctuation of the incoming traffic. However, congestion may still occur if traffic imbalance persists for some period of time.

3.1.4 Applicability to Reference Network Architecture

Any one of the contention-free switch architectures presented above can be employed for implementing a destination directed packet switch for the reference network architcture. A 590-Mbit/s throughput for packet switched traffic requires a 32-bit parallel data bus operating at 18.4 MHz or a single high-speed optic ring operating at about 600 Mbit/s (including frame overhead). Implementation of such a contention-free switch is well within the current technology.

3.2 Contention Switch Architectures

Allowing some output contention to occur in the switch (or even internal blocking within the switching fabric) can reduce the hardware complexity and speed requirement compared with the contention-free switch. From the switch capacity and hardware complexity point of view, the banyan-based switching fabric becomes the most attractive candidate for the contention switch. To resolve the output contention problem and/or the internal blocking problem, packet transfer at the input ports has to be scheduled. In each packet transfer process, a set of non-contending packets is chosen from the input ports. The packets presented to the switching fabric have distinct destination addresses and will not be collided in the switching fabric. Based on the output contention resolution scheme (or packet transfer scheduling algorithm), the contention switch architectures can be categorized into three subclasses.

The first subclass uses an output port reservation scheme at the input ports to resolve output contention. The prerequisites for this class of switches are: the switch incorporates queueing at input ports and the switching fabric is nonblocking. The function of the output port reservation scheme is to choose a nonblocking set (or a permutation set) of connections from the packets at the input ports. Due to head-of-line (HOL) blocking at the input port queue, the packet switch throughput for point-to-point connections cannot exceed 58% for a large N [5]. The throughput is defined as the average number of packets arrived to the output ports in one link slot divided by the switch size, where a link slot is defined as (packet size/input link speed). This blocking is a side effect resulting from output contention. Assume that one packet at the head of a queue cannot be transmitted due to output contention. Then, this blocked packet hinders the delivery of the next packet in the queue due to the first come first serve (FCFS) nature of the queue, even though the next packet can be transmitted to the destination without any blocking. To improve the throughput of the switch, there are three basic methods. The first method is to increase the switch speed so that more than one packets can arrive at one output port within one slot time. The ratio of the switch speed to the link speed is defined as the speedup factor (S). The second method is to use p parallel switches, p transmitters at the input port, and p receivers at the output port. The result is there are p disjoint paths between each input and output pair, the input can transmit up to p packets, and the output port can receive up to p packets at the same time. The third method is to design a more efficient scheduling algorithm to increase the throughput of the switch. In the first two methods, since more than one packet can arrive at one output port in one link slot time, the switch has to incorporate output queueing to hold the packets. In this case, each output port performs as a statistical multiplexer. Since output queueing is used, the throughput definition is modified as the average number of packets leaving the output ports in one link slot divided by the switch size.

The second subclass of switches uses the path setup strategy to resolve internal blocking of the switching fabric and output contention at the same time. The prerequisite is that the switch incorporates queueing at input ports. Due to head-of-line

(HOL) blocking at the input port queue and switching fabric blocking, the packet switch throughput is much less than 58% [6]. There are two ways to improve the throughput of the switch. The first one is to increase the switch speed so that more than one packet can arrive at an output port in one slot time. The second one to use *p* parallel switches, *p* transmitters at the input port, and *p* receivers at the output port. This will yield *p* disjoint paths between each input and output pair so that more than one packet can arrive at an output port at the same time, and the input port can try to transmit up to *p* packets at the same time. Since more than one packet can arrive at one output port within the same slot time, the switch has to incorporate output port queueing to hold the packets.

The third subclass of switches uses an address filter to select the packets destined to the output port without any output reservation or path setup scheme. The prerequisite is that a packet can reach the destined output port without any blocking and output contention. A switch provides a disjoint path between each input and output pair. Since there is a disjoint path between each input and output pair, the switching fabric is point-to-point nonblocking and point-to-multipoint nonblocking. However, since the format of the point-to-point routing tag is different from that of the point-to-multipoint routing tag, the implementation of the switch such as the address filter design or the switching element design is different for point-to-point and multicast cases even though the switching architectures remain the same.

3.2.1 Output Port Reservation Scheme

The output contention problem is resolved using the output port reservation scheme at the input ports. Since there is no internal blocking for this class of switches, if the output port of a packet is reserved, the path through the switch is also reserved.

Among point-to-point nonblocking switching fabrics based on the banyan network, the sorted-banyan-based network is the most widely used network. Before the sorted-banyan-based switch is described, the two essential components, i.e., the banyan network and the batcher sorting network, are introduced.

A banyan network is in the category of multistage interconnection networks [7]. It can be constructed using any size of switching elements. If the size of the switching elements in the banyan network is a D x D switching element, the number of switching elements at each stages is N/D, and the number of stages is Log_D N. The banyan network is a unique path network in which there is only one path between any input-output pair. The banyan network is topologically equivalent to many other multistage interconnection networks such as baseline, omega, flip and shuffle networks.

A 2 x 2 switching element has four allowed states: straight, exchange, lower broadcast, and upper broadcast. For the point-to-point banyan network, only the straight and exchange states are used, and each switching element needs to check only one bit of the routing tag to route the packet. The lower broadcast and upper broadcast states are the basic principles that a banyan network can perform the multicast function; the multicast banyan network will be discussed in a latter subsection. If the corresponding routing bit is zero, the data will be sent to the upper link of the element; otherwise, to the lower link. For easy hardware implementation, the switching element at stage 1 checks bit 1 of the routing tag. The switching element at stage k checks bit k of the routing tag, where $1 \leq k \leq \mathrm{Log}_2 N$. Following this bit representation, the leftmost bit of the routing tag is the least significant bit and the rightmost bit is the most significant bit.

The batcher sorting network is in the category of bitonic sorting networks which produce sorted outputs from circular bitonic inputs [8]. A bitonic list is a list which monotonically increases from the beginning to the i-th element and then monotonically decreases from the i-th element to the end. A circular bitonic list is created from joining the beginning and the end of a bitonic list, and then breaking the circular structure into a linear structure at any desired point.

The sorting network has a similar property as a banyan network, i.e., a large network is constructed recursively from a smaller network. An N x N batcher sorting network has $\frac{1}{2} \mathrm{Log}_2 N (\mathrm{Log}_2 N + 1)$ stages, and each stages consists of $\frac{N}{2}$ sorting elements.

One of the important properties of the banyan network is that if the incoming packets are arranged either in ascending or descending orders and there is no empty line between any two active lines, there is no internal blocking within the banyan network. An active line means that there is a packet waiting to be transmitted. A' way of arranging the arriving packets in a descending order is to use a batcher sorting network. To concentrate the packets at the outputs of the sorting network, empty packets are generated at the inactive input ports. The result is that the total number of packets (data packets and empty packets) generated at the input ports is always equal to the size of the switch. To accommodate the empty packets within the sorting network, the routing tags are modified as follows. The most significant bit of the routing tag is designated to be the activity bit. For data packets, the activity bit is 1; for the empty packet, the activity bit is 0. After the sorting network, the appearance of the data packets at the outputs of the sorting network will be above the empty packets. Hence, the data packets have been concentrated and arranged in a descending order at the outputs of the sorting network. The empty packets are deleted and the data packets are fed into the banyan network. The sorting network and the banyan network are connected using a shuffle interconnection pattern (see Figure 8).

3.2.1.1 Point-to-Point Sorted-Banyan-Based Switch

Since the sorted-banyan-based switch has a point-to-point nonblocking switching fabric, if the destined output port of a packet is reserved, the path through the switch is also reserved. The possible output contention resolution schemes are as follows:

- output reservation at the input ports with input buffering
- setup phase + transfer phase with input buffering
- transfer phase + retransmission phase with input buffering

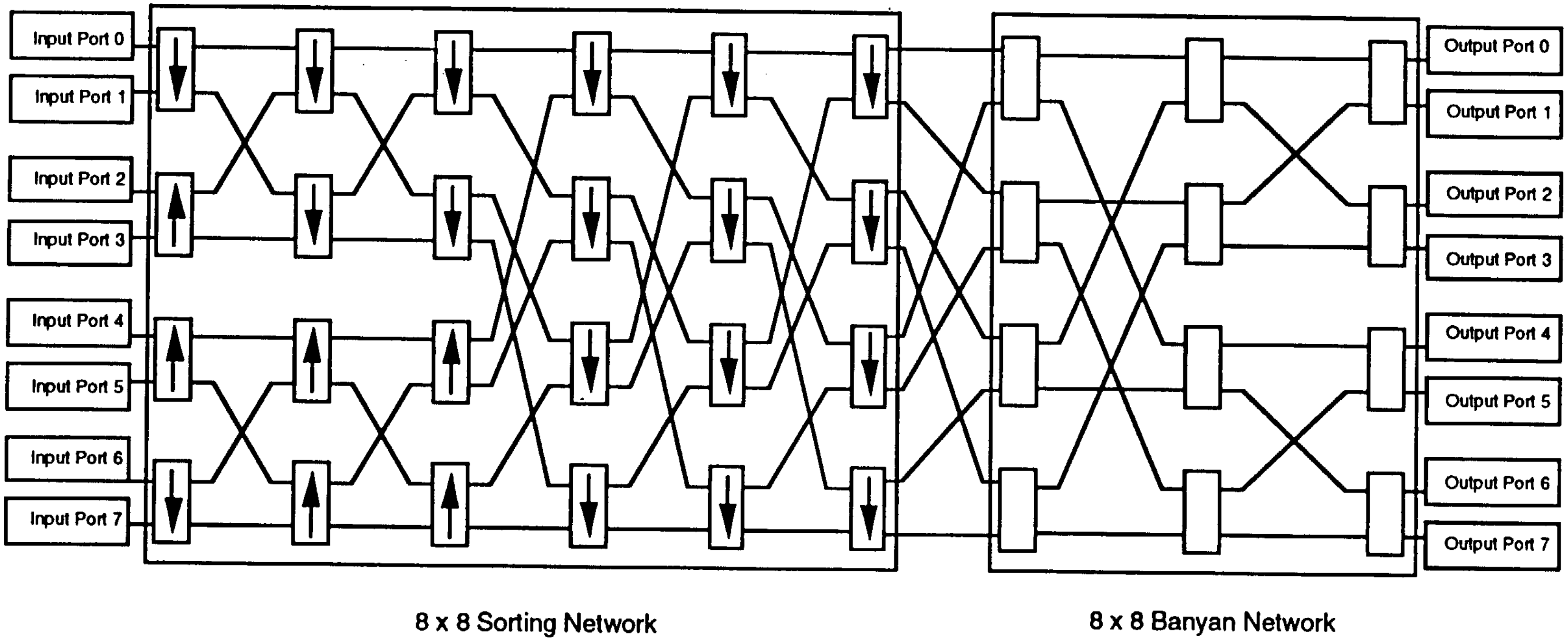


Figure 8. Sorted-Banyan-Based Network (Point-to-Point Nonblocking Network)

- trap phase at the sorted-banyan-based switching fabric with input buffering
- sorted-banyan-based switching fabric with three-phase algorithm

a. Output Reservation Scheme 1: Ring Reservation Scheme

Basically, the ring reservation scheme uses the token ring principle to resolve output contention [9]. The input ring connects all the input ports of the switch, and a stream of tokens, where one token represents one output port, are sent through the input ports (see Figure 9). The function of the ring is to perform output reservation for each input port. At the beginning of every slot time, the ring module sends a stream of tokens and passes these tokens to all the input ports. The input port searches the right token according to the destination routing tag of the current packet. If the token for the corresponding routing tag is on the stream, then the token is removed so that no other input port can transmit a packet to the same output port at the same slot time. After the token stream has passed through all the input ports, the input ports that have reserved a token can transmit the packet at the beginning of the next slot time. In implementation, only one bit is necessary for one token. For example, value 1 represents there is a token and value 0 no token. To assure fairness among the input ports of accessing the tokens, several strategies can be considered. The first is at different slot time, the stream will be started at different input port. The second is to send this stream from the beginning of the input ports and then from the end of the input ports alternatively.

To improve the throughput of the switch, a non-FIFO input queue with the windowing scheme is used. In this scheme, if the first packet is blocked due to output blocking, the scheduling algorithm also examines (searches) the packets on the back of the first packet. This scheme is also referred to input queue by-pass [10]. The number of packets examined each time depends on the preset window size or the checking depth. If one of the packets within the checking depth has a chance to be transmitted, this packet will be transmitted first. In this sense, the FCFS input queue has a checking depth 1 while a non-FIFO input queue has a checking depth greater than 1. Theoretically, if the checking depth is infinite, the throughput of the switch can reach 1. However, in practical, the checking depth is finite and less than O (10). The effectiveness of using a non-FIFO queue with a finite checking depth is examined in a latter subsection using simulation techniques.

b. Output Reservation Scheme 2: Output port Reservation with an Arbitrator

The other way of resolving the output contention problem is to use a bidirectional sorting network [11] (see Figure 10). The sorting network has the property of arranging the arriving packets based on their destination addresses either in the ascending or descending order. At the beginning of every slot, the input ports send setup packets containing the destination addresses (or routing tag) to the bidirectional sorting network, and finally reaches the arbitrator. All the setup packets that have the same destination addresses will be adjacent to each other. A distinctive set of destination

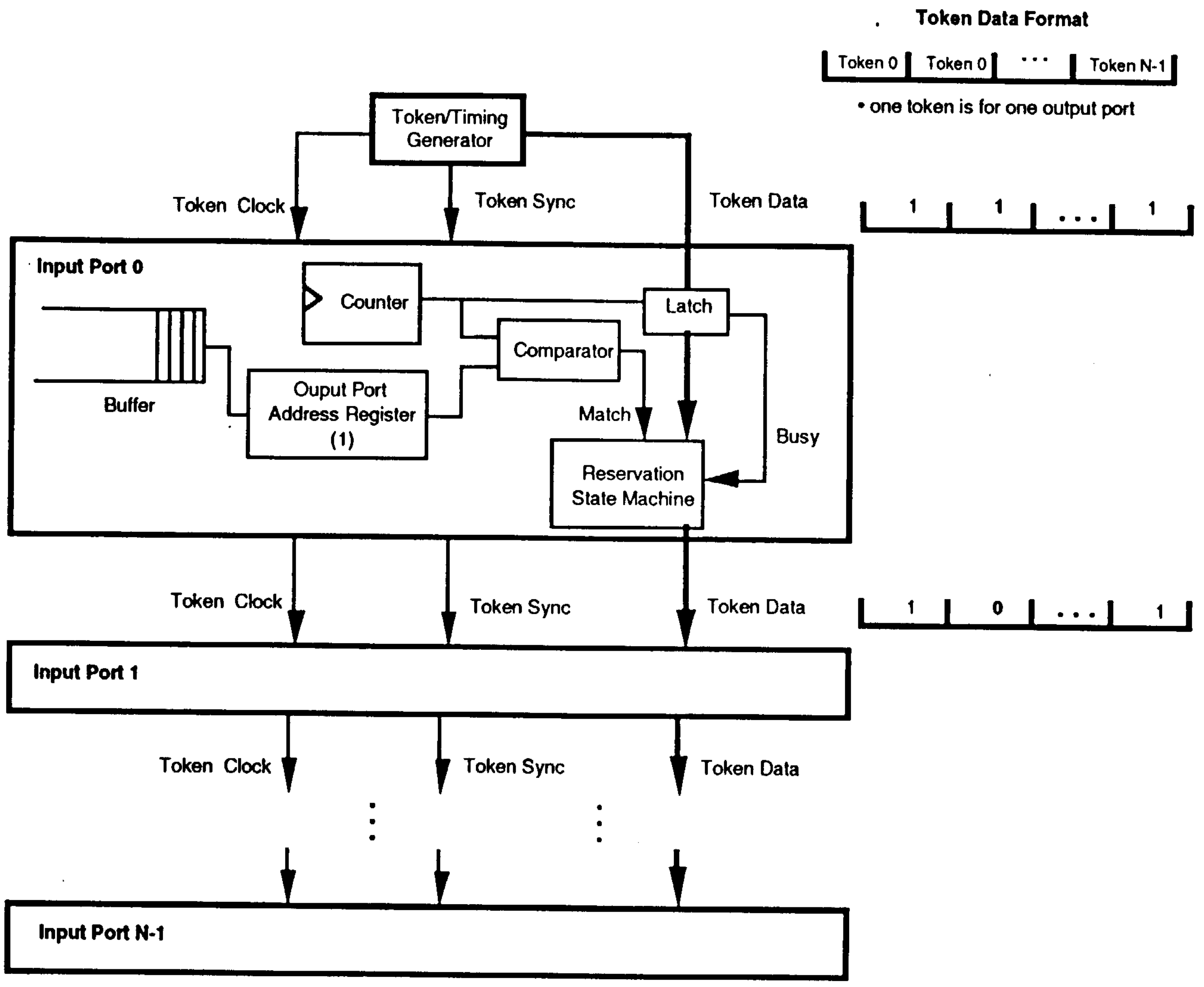


Figure 9. Output Port Reservation Scheme with Tokens

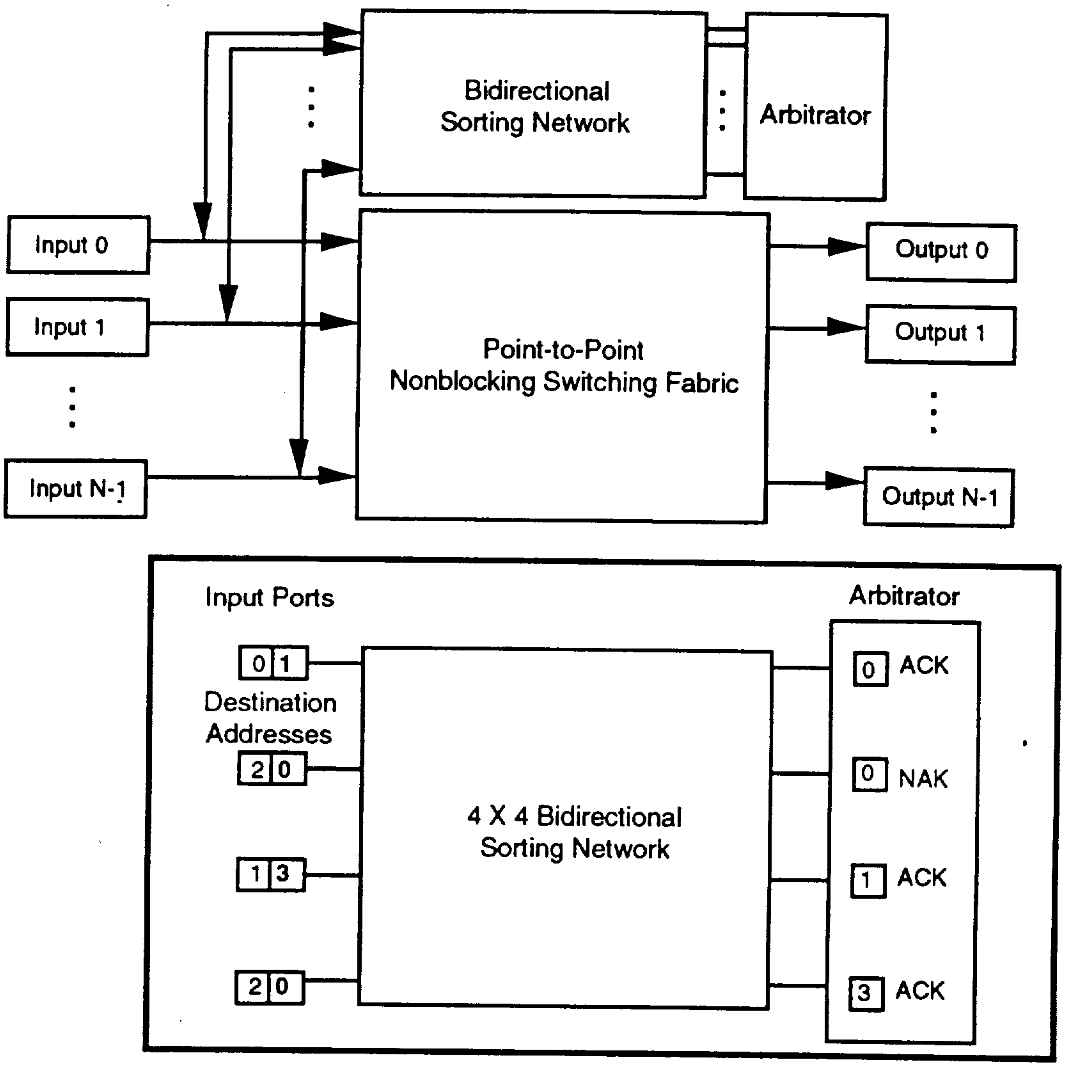


Figure 10. Output Contention Resolution Using a Sorting Network

addresses can be selected easily using an array of comparators within the arbitrator. The arbitrator sends acknowledgements (ACKs) and negative acknowledgements (NACKs) through the bidirectional sorting network to the input ports to report the arbitration result. For the input ports whose packets have been selected for transmission at the next slot time receive an ACK. For the input ports whose packets have not been selected for transmission at the next slot time receive an NACK.

To improve the throughput of a switch, a large checking depth is desirable, where a throughput is the average number of packets arrived at the output ports in one link slot time. To examine more depths, the input port which has received an ACK will send the same routing tag again. To guarantee that the packets which have won the arbitration at the previous run still win the arbitration at this run, the routing tags of these packets will be prepended with a priority bit so that these routing tags always win the arbitration. The input port which has received an NACK will send the routing tag of the packet behind the HOL packet. With this priority bit mechanism, the input port is able to check more depths to improve the switch throughput.

c. Output Reservation Scheme 3: an Output Contention Resolution Device

The output reservation is accomplished using an output contention resolution device [12]. Within this device, there are two arrays of registers A and B, where the number of registers in each array is N and the size of register is the size of the routing tag that are used to hold the routing tags from all the input ports (see Figure 11). There is another array of bit registers R to hold the reservation result. If R_i is 0 at the end of output port reservation process, then input port i can transmit the current packet in the next slot. Initially, all the routing tags from the input ports will be loaded into the array A and array B; hence, the contents of array A and array B are exactly the same. All the bits in array R are 0. To reserve the output ports, the routing tags between A and B have to be compared with each other so that a distinctive set of routing tags can be selected. This operation is achieved by fixing array A and rotating array B. After each rotation, the contents of array A and array B are compared. If $A_i \neq B_i$, there is no action. If $A_i = B_i$, then one routing tag will be selected for transmission. Now the problem is which routing tag should be selected. To resolve this problem, another array of priority bit registers, P, is used.

Initially, all the bits in array P are 0. Starting from the first rotation cycle, a bit 1 is loaded into P_0 (see Figure 12). In this situation A_0 has the routing tag from input port 0 and B_0 has the routing tag from input port N-1. If $A_0 = B_0$ and now $P_0 = 1$, R_0 remains 0. Thus, the routing tag at A_0 wins the arbitration. A_1 has the routing tag from input port 1 and B_1 has the routing tag from input port 0. If $A_1 = B_1$ and $P_1 = 0$, R_1 is changed to 1. This means the routing tag at A_1 loses the arbitration. It can be seen that if $P_i = 1$, it means that the input port number at B_i is larger than the input port number at A_i. If $P_i = 0$, it means that the input port number at B_i is smaller than the input port number at A_i. It can be observed the arbitration rule for $A_i = B_i$ situation is that whoever holds the routing tag from an input port of a smaller number wins the arbitration. This means the priority is given from top to down of the input ports. This priority is implemented using the priority bit register P.

At the second rotation cycle, P_0 and P_1 all have bit 1. The comparison is performed between array A and array B following the same procedure mentioned above.

In summary,

- after every rotation, the contents of A_i and B_i are compared.
  - if $A_i \neq B_i$, no action.
  - if $A_i = B_i$,

    if $P_i = 0$, $R_i = 1$

    if $P_i = 1$, no action.

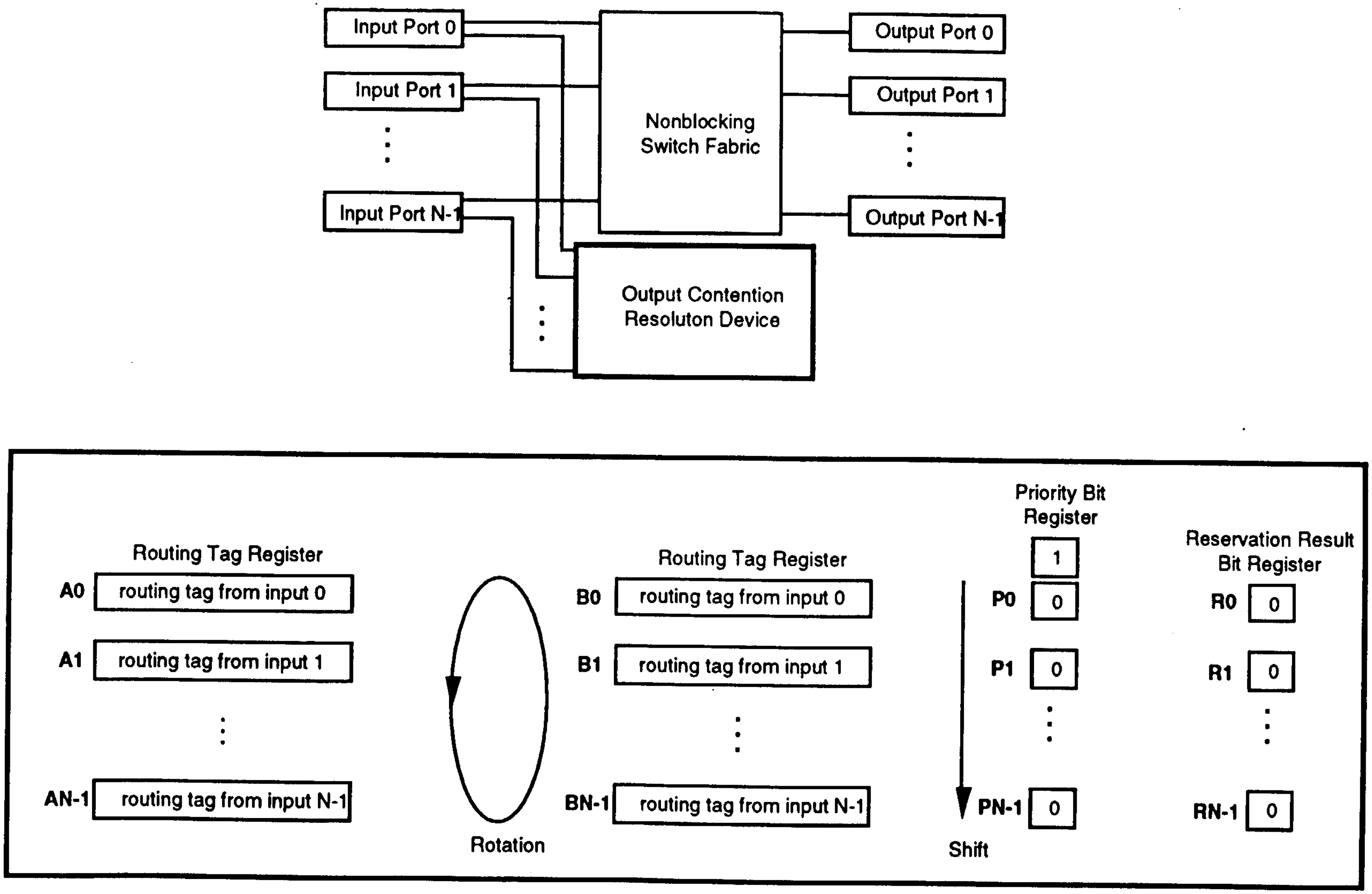


Figure 11. Output Contention Resolution Device

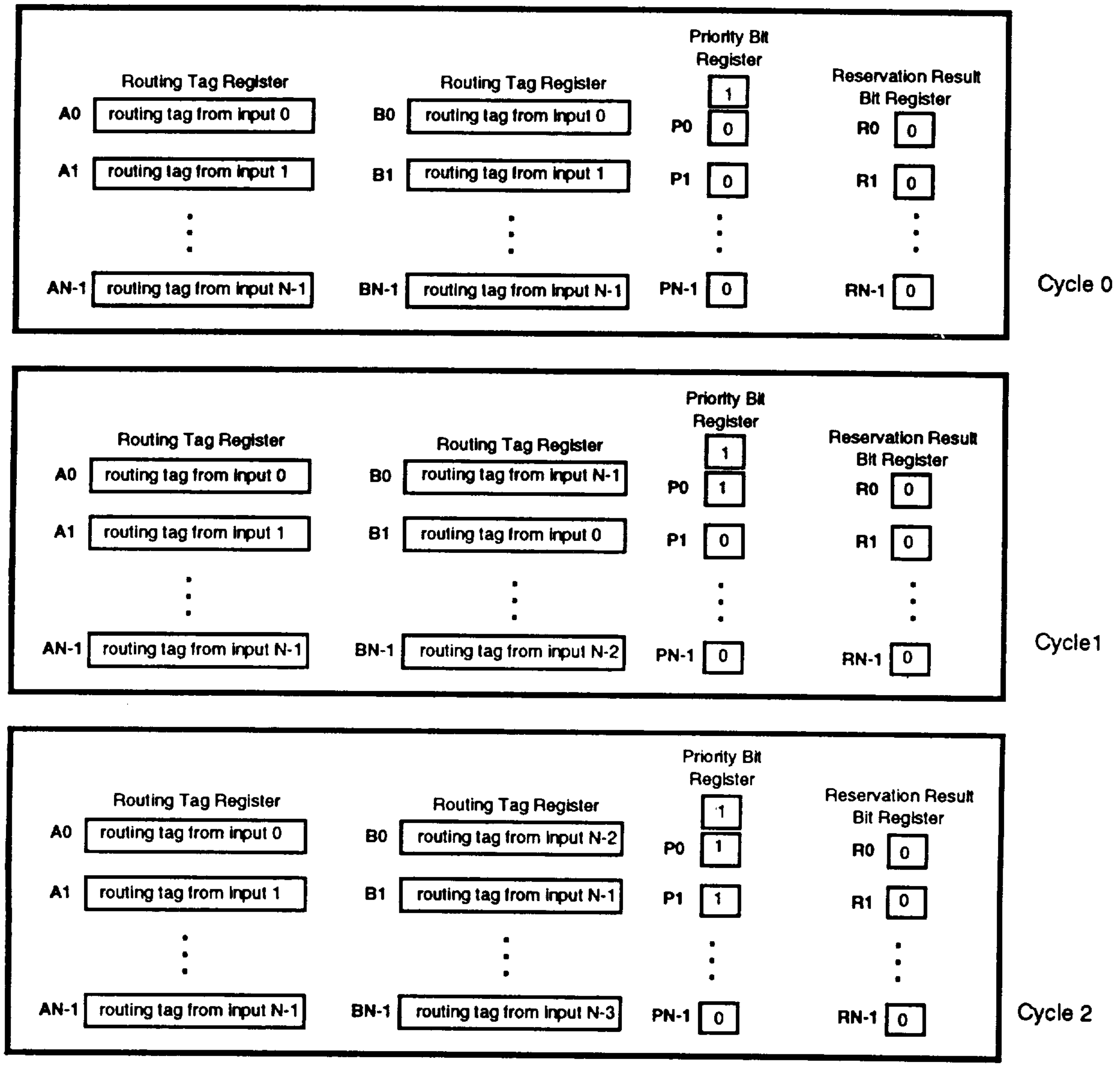

Routing Tag Register
A0 routing tag from input 0
A1 routing tag from input 1
AN-1 routing tag from input N-1
Routing Tag Register
B0 routing tag from input 0
B1 routing tag from input 1
BN-1 routing tag from input N-1
Priority Bit Register
1
P0 0
P1 0
PN-1 0
Reservation Result Bit Register
R0 0
R1 0
RN-1 0
Cycle 0
Routing Tag Register
A0 routing tag from input 0
A1 routing tag from input 1
AN-1 routing tag from input N-1
Routing Tag Register
B0 routing tag from input N-1
B1 routing tag from input 0
BN-1 routing tag from input N-2
Priority Bit Register
1
P0 1
P1 0
PN-1 0
Reservation Result Bit Register
R0 0
R1 0
RN-1 0
Cycle1
Routing Tag Register
A0 routing tag from input 0
A1 routing tag from input 1
AN-1 routing tag from input N-1
Routing Tag Register
B0 routing tag from input N-2
B1 routing tag from input N-1
BN-1 routing tag from input N-3
Priority Bit Register
1
P0 1
P1 1
PN-1 0
Reservation Result Bit Register
R0 0
R1 0
RN-1 0
Cycle 2

Figure 12. Output Contention Resolution Device

- after N-1 rotations, all the input port i with $R_i = 0$ can be transmitted.

To have a fair access to an output port, the priority bit in P_0 can be loaded to different P_i at the beginning of arbitration at different slots. To check more depths into the input buffer, another array of bit registers W is required. Bit registers W are used to record the arbitration results of the previous runs. Initially, all the bits in the array W are 1. At the end of the first-run arbitration, the results of array R will be copied into array W. At the second run of the arbitration, the packets which lost the first run arbitration will send the routing tags of the packets behind the HOL packets. The packets which won the first run of arbitration will send the same routing tags as the first run. During the arbitration, if $W_i = 0$, R_i will be kept the same as the first run. This means that the packets which won the first run of arbitration are guaranteed to win the second run of arbitration.

d. Output Reservation Scheme 4: Setup Phase + Transfer Phase Protocol

The procedure of this protocol is shown in Figure 13. The input port sends a small setup packet and attempts to reserve a path between the input port and the destined output port. The setup packet consists of only the routing tag. If the output port receives the setup packet, the output port sends an acknowledgement (ACK) back to the originating input port. After the path has been successfully set up, the input port can release the packet and send it to the output port. If the input port does not receive an ACK within three routing tag's unit time (two tag's time for the round trip delay time and one tag's unit time for the transmission time), then the input port sends the setup packet again, and the whole procedure is repeated. From the above discussion, the switch needs to have bidirectional connection capability. This method can be operated in the minislot mode, where the length of the minislot is the setup time (three routing tag's unit time). Note this method can also be used for a blocking switching fabric. This issue will be discussed in detail in a latter section.

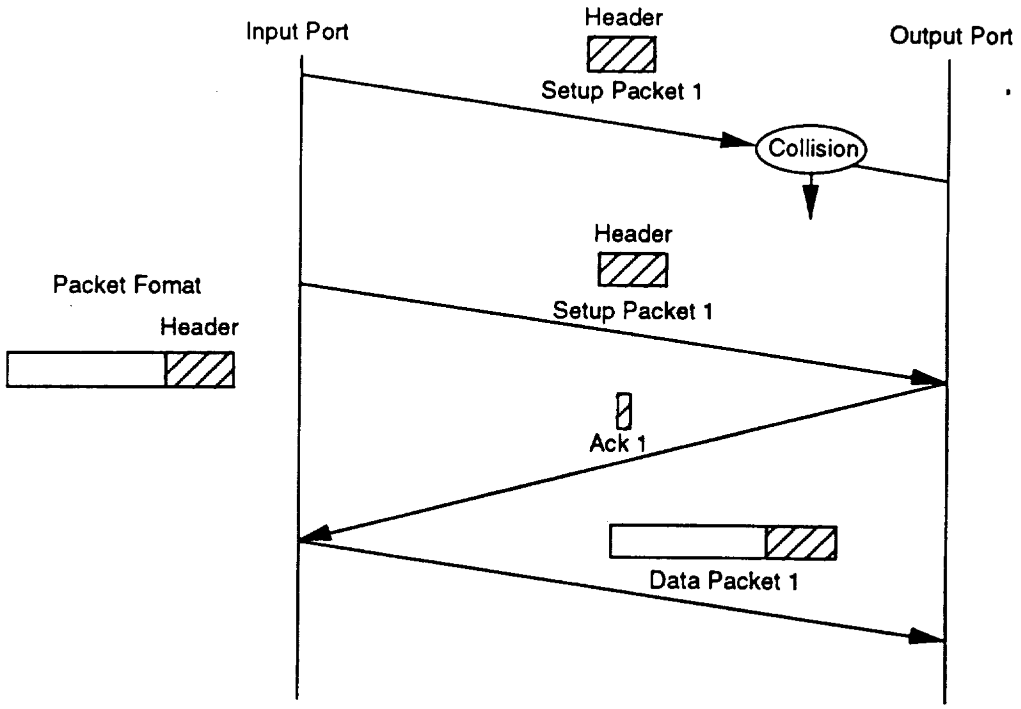


Figure 13. Setup Phase + Forwarding Phase Protocol

e. Output Reservation Scheme 5: Transfer Phase + Retransmission Phase

The procedure of this protocol is shown in Figure 14. This procedure is only suitable for the slotted mode operation. First, the input port stores a copy of the packet in the buffer. Then, the input port sends the whole packet to the destined output port. When

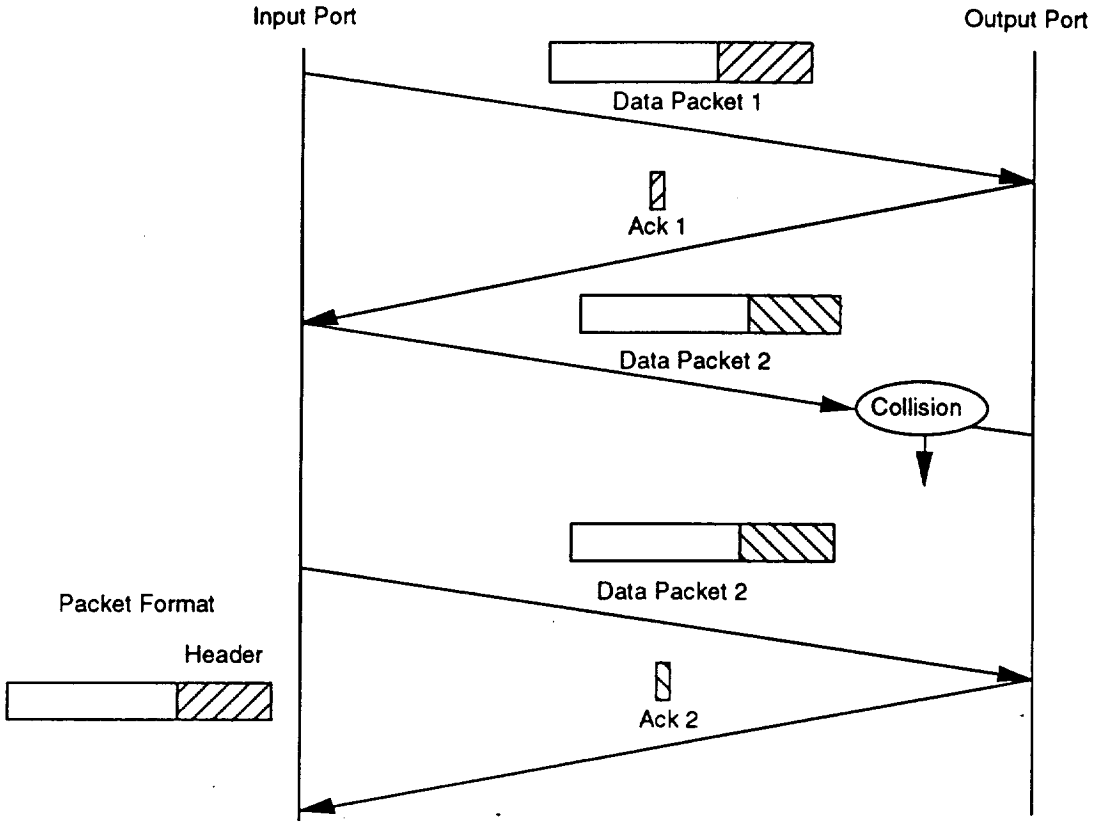
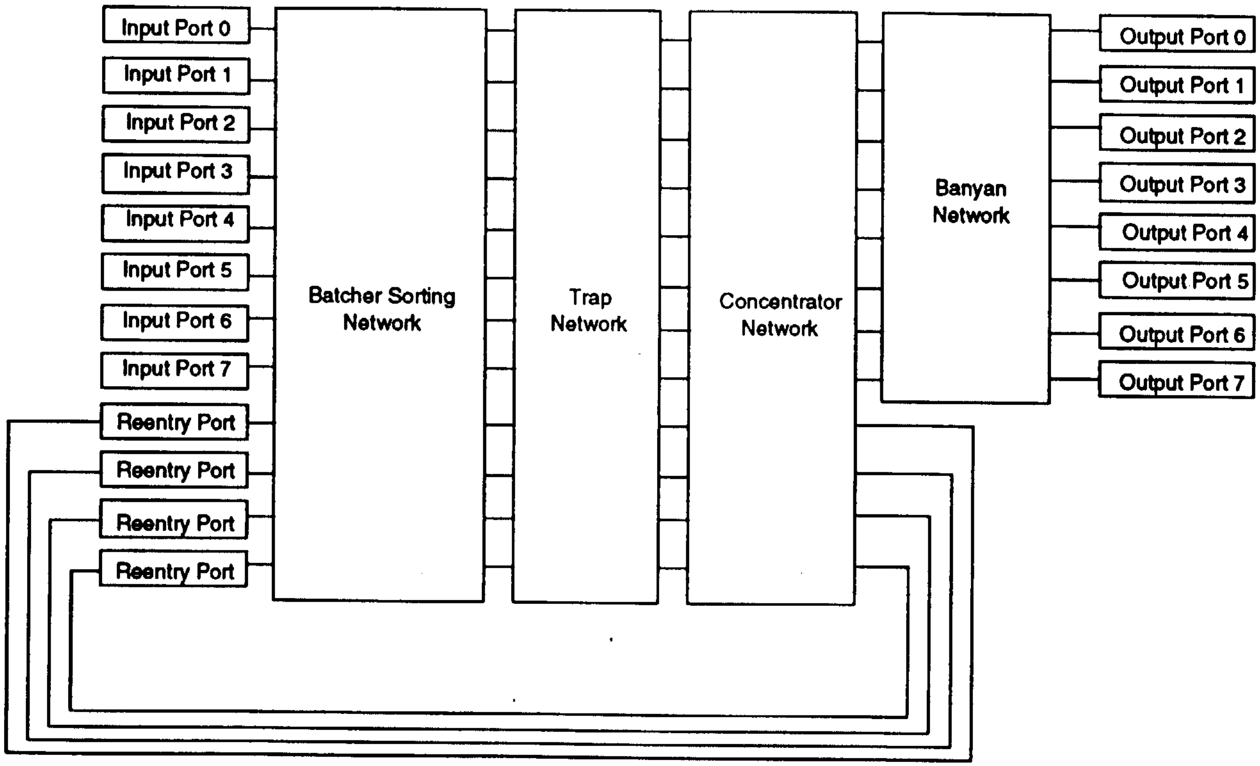


Figure 14. Forwarding Phase + Retransmission Phase Protocol

the output port receives a packet, an ACK is sent back to the originating input port. When the input port receives an ACK, the input port discards the packet and processes the next packet waiting in the buffer. If the ACK does not come back within two routing tag's time plus one packet length's time, then the packet is sent again and the whole procedure repeats. The switch needs to have bidirectional connection capability. The method can only be operated in slotted mode. The length of the slot is the packet length's time plus two routing tag's time. Note this method can also be used for a blocking switching fabric. This issue will be discussed in detail in a latter section.

f. Output Reservation Scheme 6: Trap Phase at the Sorted-Banyan-Based Switching Fabric with Reentry Network

The scheme uses a trap network after the sorting network [13] (see Figure 15). The trap network resolves output contention by marking the packets with distinct output addresses. In implementation, the trap network is implemented using an array of comparators. After the trap network, there is a concentrator. The concentrator sends the marked packets to the banyan network so that packets presented to the banyan network all have distinct destination addresses. The concentrator sends the trapped packets back to the reentry inputs of the sorting network. The packets in the reentry



- Trap network resolves output contention by marking packets with distinct output addresses

Figure 15. Sorted-Banyan-Based Network with Reentry Ports

port are retransmitted during the next time slot. The size of the sorting network is larger than the size of the switch to accommodate the reentry ports. If the number of trapped packets is larger than the number of reentry ports, the packet will be lost. Also the packets may be delivered out-of-sequence, because the trapped packets are sent back to the reentry ports (not the original input ports). These retransmitted packets have to be given a priority higher than that of the new packets when conflict occurs at the output port; otherwise, there are chances that packets are transmitted out of sequence.

g. *Output Reservation Scheme 7: Sorted-Banyan-Based Switching Fabric with 3-Phase Algorithm*

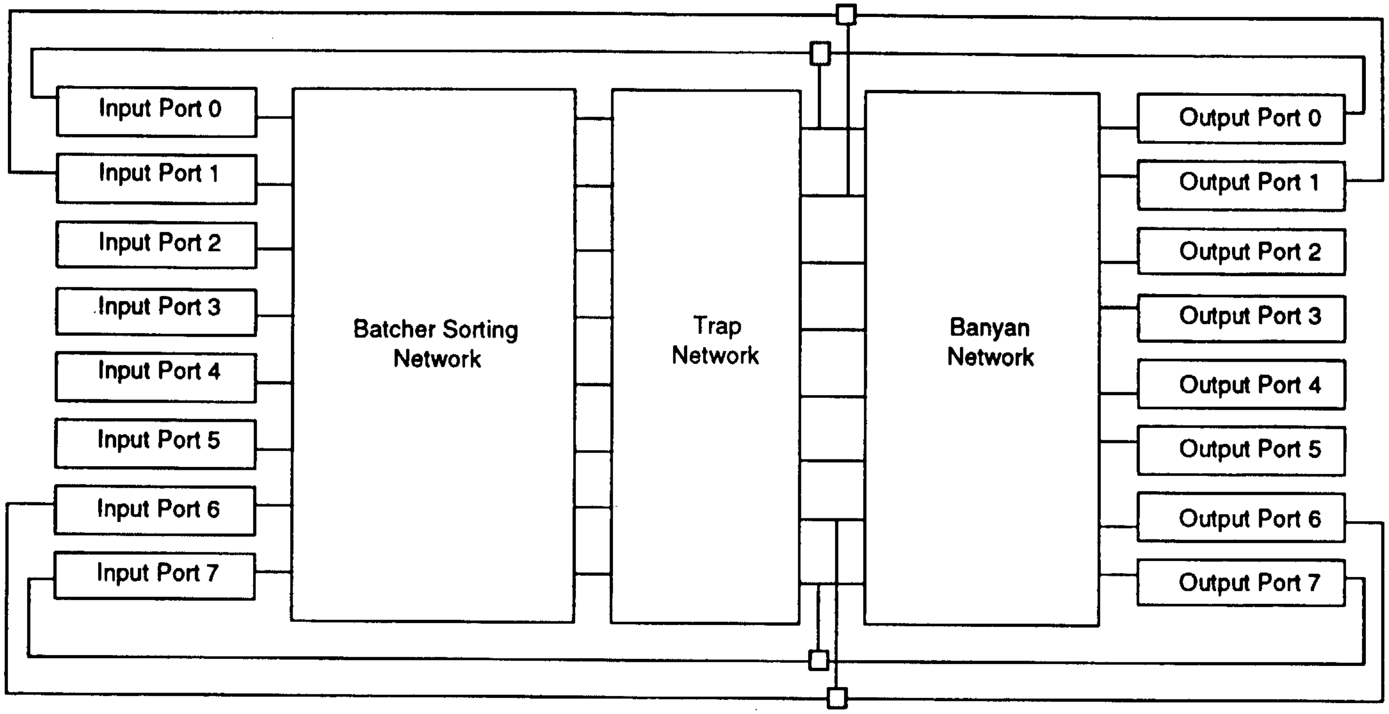
The output contention resolution algorithm is divided into three phases [14] (see Figure 16). At Phase 1 the input ports send setup packets to the trap network to resolve output contention, where the setup packet contains the source address and the destination address. At Phase 2 the trap network marks the setup packets with distinct destination addresses. An ACK packet will be sent back to the originating input ports for the marked setup packets, where the ACK packet contains the source address. To achieve this function, the outputs of the trap network and the corresponding input ports are connected. All the ACKs are sent to the input ports from the trap network first. The sorted-banyan-based network routes these ACK packets using the source addresses to the corresponding output ports. The input port and the corresponding output port are also connected together. Hence these ACK packets are sent from the output ports to the corresponding input ports. At Phase 3 the input ports that receive ACK packets send the data packets prepended with the destination routing tags to the output ports.

Congestion Issue

For the switches mentioned above, the possible situations which may cause congestion at the input ports of the switch are:

- burst arrivals of packet destined to the same output port (or the downlink beam).
- nonuniform output destination distribution of the traffic.
- nonuniform traffic intensity among the input ports.

To tackle the congestion problem, a congestion avoidance technique has to be employed. This can be performed by monitoring the downlink beam utilization and the input buffer queue length. The information is broadcasted back to the earth stations continuously. If the utilization and/or the queue length exceeds a certain threshold, the earth station will defer sending the packets destined to the congested downlink beams.



- Phase 1: Input ports send setup packet (source address + destination address) to the trap network
- Phase 2: Trap network resolves output contention by sending ack packets with distinct destination addressess back to the input ports and these ack packets will be routed to the original input ports using the source addresses
- Phase 3: The acknowledged input ports send the data packets to the output ports.

Figure 16. Sorted-Banyan-Based Network with 3-Phase Algorithm

If input and output buffering are used at the same time, accumulation of packets occurs either at the input or at the output. Hence, it is possible to reduce the congestion by shifting traffic to the uncongested port. The shifting effect allows the congested port to digest the traffic and return to the normal state while the uncongested port tries to absorb the excessive traffic. This is to say that by utilizing the buffer space intelligently, congestion may be reduced to a minimum.

3.2.1.2 High Speed Bus with Distributed Input Memories

As mentioned above, the TDM bus is a degeneration of the banyan switch. In this scheme, the buffering of the arriving packets is performed in the input ports. As shown in Figure 17, the distributed input memory approach is suitable for consistent frame format between input lines and output lines. Since there is no output buffer, the output contention has to be resolved at the input port. Hence, an output port reservation device such as the ring reservation module is necessary to schedule the packet transmission sequence among the input ports.

3.2.1.3 Contention-Free Switch

It is possible to create a contention-free banyan-based switch. A contention-free switch is defined as a switch whose output port can receive up to N packets in one link slot time, where N is the size of the switch. If the switch speed is increased to N times of the link speed, then the output port can receive up to N packets in one link slot time. If there is a disjoint path between any input and output pair and there are N receivers at the output port, then the output port can receive up to N packets in one link slot time. A parallel switch consisting of N nonblocking banyan switches is contention free. Two examples are given below. In these examples, only output buffering is required since the switch itself is contention free.

a. Contention-Free Sorted-Banyan-Based Switch

To design a contention-free switch based on the sorted-banyan-based switch is to operate the switch N times faster than the link speed, where the N is the size of the switch. Evidently, this method is not useful if the link speed is already high or the switch size is large.

b. Contention-Free Parallel Switch Architectures

To have a contention-free switch, the number of switching fabrics stacked in parallel and the number of receivers at the output ports have to be the same as the switch size. It is possible to use only one switching fabric to construct a contention-free switch. However, the switching fabric becomes nonsymmetric, i.e., the number of outputs is larger than the number of inputs. To have a contention-free switch, the switching fabric size has to be N x N^2. One output port is interfaced with N outputs of the switching fabric (see Figure 18). Since more than one packet can arrive at one output port at the same time, output queueing is necessary to hold the packets.

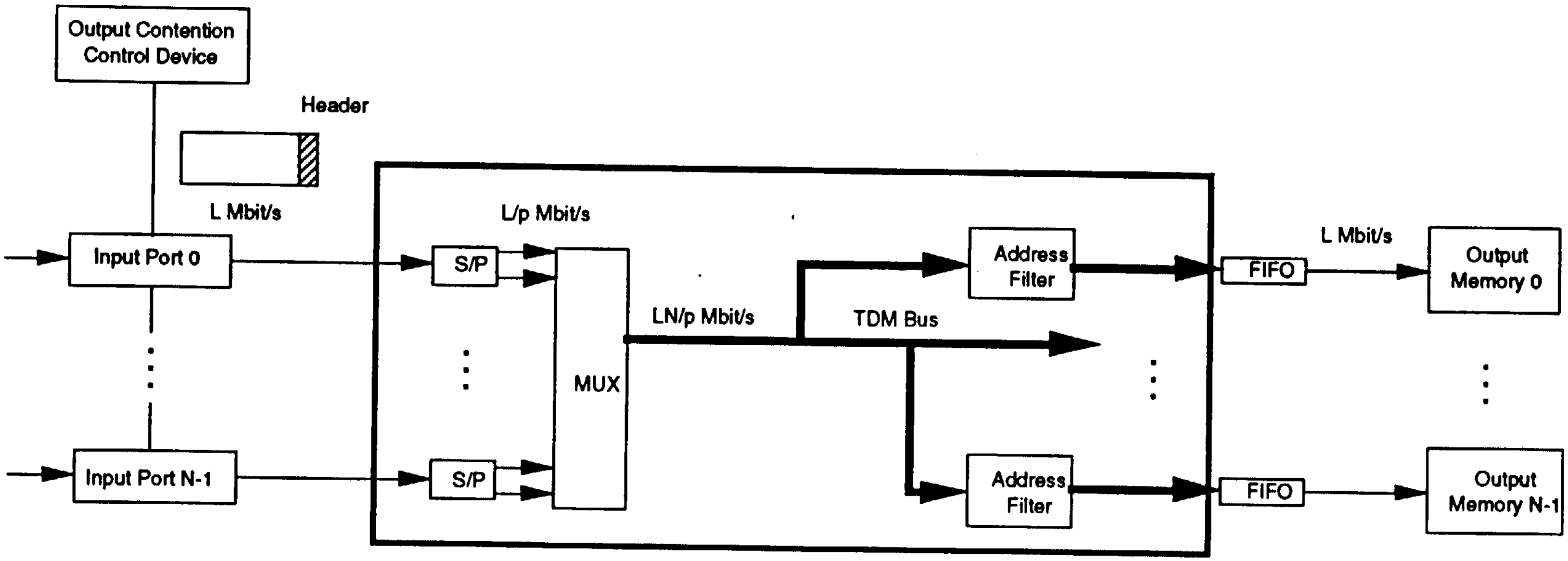


Figure 17. TDM Bus with Distributed Input Memories

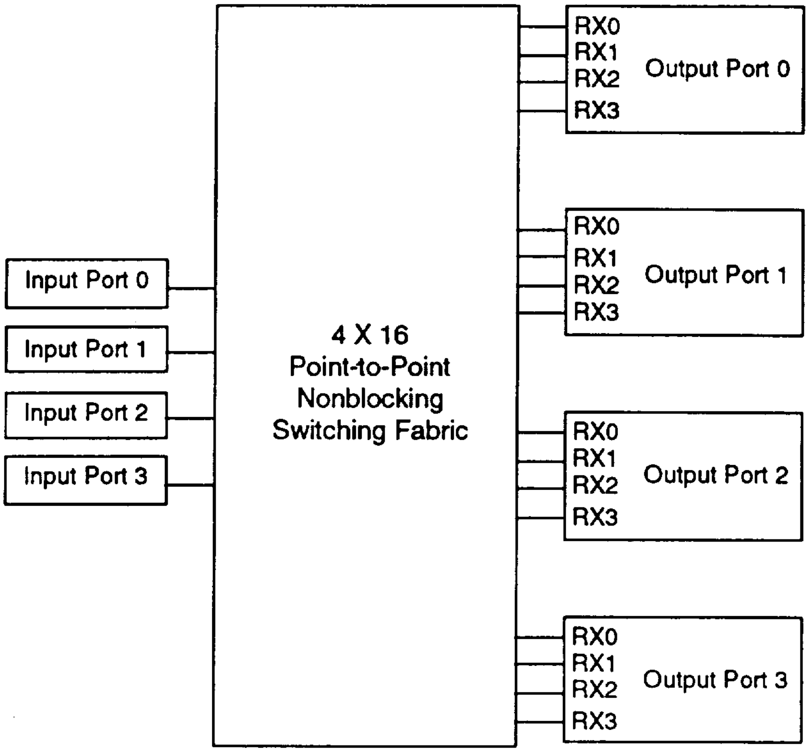


Figure 18. 4 x 4 Contention-Free Switch

Congestion Issue

In the above two switching architectures, there is no output contention. However congestion may occur at the output ports. The congestion situation is similar to the TDM bus with output memories. It will not be repeated here.

3.2.1.4 Multicast Unbuffered Banyan Switches

There are three configurations of multicast switches depending on where the multicast packet is duplicated. The first one duplicates the multicast packet at the input port one by one, i.e., using the store-and-forward at the input port approach. The second one duplicates the packet at the switching fabric, i.e., the point-to-multipoint switching fabric approach. The third one duplicates the packet at the output port, i.e., the multicast modules at the output ports. Note that if the switching fabric is nonblocking, the output reservation schemes used for point-to-point connections can also be used for point-to-multipoint connections with a slight modification. The output reservation schemes for point-to-point connections assume that each input port can reserve one output port at a time. For point-to-multipoint connections, each input port can reserve more than one and up to N output ports at a time.

These configurations are summarized below.

- Store-and-Forward at the Input Port
  - point-to-point nonblocking switching fabric.
  - packet duplication occurs at the input port.
- Sorted-Multicast-Banyan-Based
  - point-to-multipoint nonblocking switching fabric.
  - packet duplication occurs at the switching fabric.
- Multicast Modules at the Output Port
  - one point-to-point nonblocking switching fabric for point-to-point connections.
  - one point-to-multipoint nonblocking switching fabric at the output ports for point-to-multipoint connections.

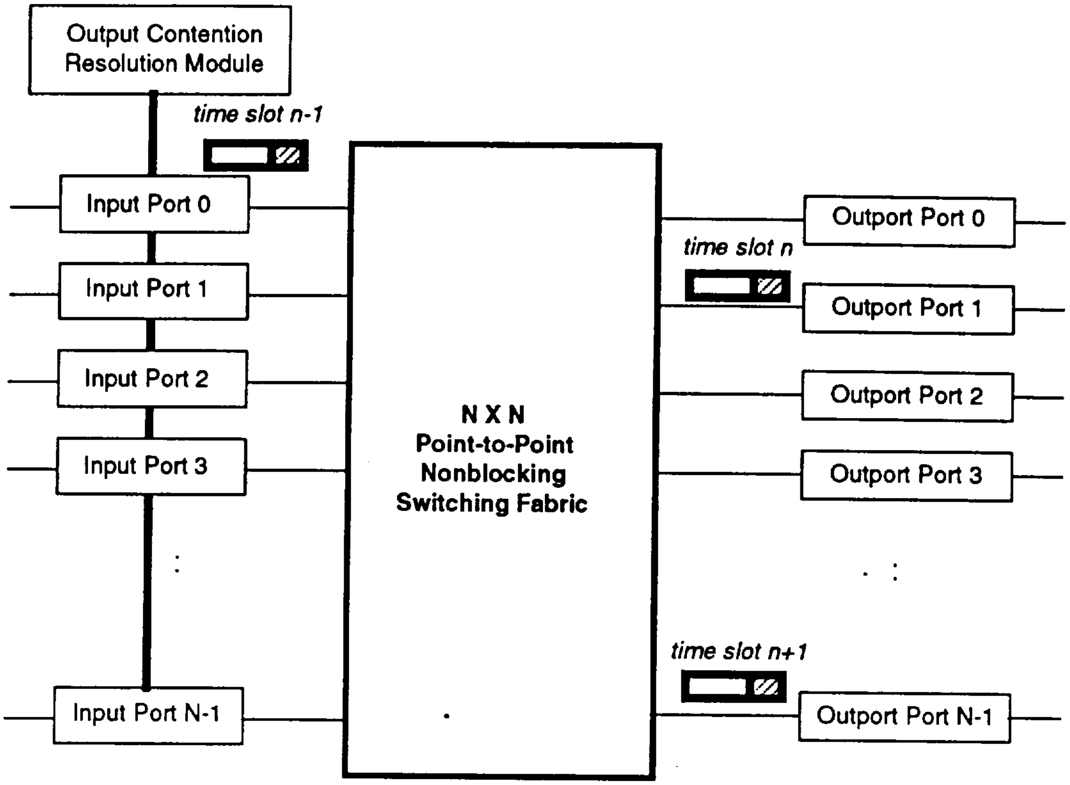
a. Store-and-Forward at the Input Port

In this approach, the multicast operation is achieved by sending the multicast packet one by one from the input port (see Figure 19). The advantage of this approach is that a point-to-point switch can be used as a multicast switch; hence, the hardware cost for building a multicast switch is minimal. The disadvantages of this approach are the long delay due to the serial transfer of the multicast packet and serious congestion if the number of duplication is large.

The above approach is feasible and very cost effective if the amount of multicast traffic is small and the number of duplication of each multicast packet is small. Otherwise, serial packet duplication at the input port has to be modified so that parallel duplication is possible.

One of the methods for parallel duplication is to send the multicast packet to adjacent input ports so that packet duplication can be achieved in parallel by many input ports, and each input port only handles a portion of the multicast traffic. In a sense, a virtual copy network is implemented among the input ports using a bus structure. It can be envisioned that this procedure involves a lot of handshaking among different input ports.

If the input and output ports are combined into one module, the switching fabric can be used as a copy network. Hence, the input port can send the multicast packet to several output ports, and the output ports can relay this multicast packet to the accompanying input ports. The locations of the input ports which are used to duplicate the packets are decided at the call setup time.



• The multicast packet is sent to different output ports one by one from the input port.

Figure 19. Store-and-Forward at the Input Ports

Congestion Issue

Congestion occurs in this class of switches due to the following two reasons:

- traffic imbalance of point-to-point connections.
- traffic imbalance of point-to-multipoint connections.

The main reason for traffic imbalance of point-to-multipoint connections is that if the number of duplication of each packet is large, congestion occurs due to the serial transfer of the multicast packet at the input port.

One possible solution for the traffic imbalance of point-to-multipoint connections is that during the call setup phase, only a very small amount of multicast traffic can be accepted. In essence, a very conservative call admission control is applied to ensure that the multicast traffic almost never exceeds the capacity.

b. Sorted-Multicast-Banyan-Based

The switching fabric is based on the multicast banyan network. As in the point-to-point banyan network, the multicast banyan network has internal blocking. It is found that the multicast banyan network can become a nonblocking multicast switching network by using a sorting network in front of every stage of the multicast banyan network [15].

Input buffering is used to hold the arriving packets. It is assumed that the input port has the call splitting capability such that the transfer of the packet can be partially completed. To have a consistent operation of the switching network, empty packets are generated at the input ports if no packets are ready to be transmitted at a slot time so that the total number of packets at the switching network is always equal to the size of the switch.

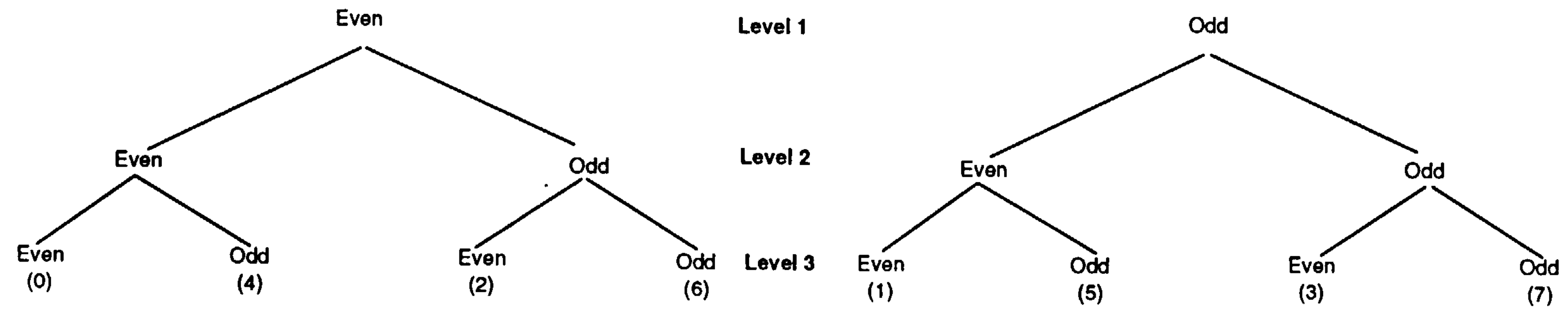
The multicast routing field formats use the even and odd group concept associated with the levels of the switching network, and they are arranged using a tree hierarchy structure (see Figure 20). The definition of a level in the proposed switching network will be explained later. At level 1, the even group consists of the output addresses whose modulo 2 results are 0; the odd group consists of the output addresses whose modulo 2 results are 1. The addresses at level 1 consist of 2 bits which are used for routing at level 1 of the switching network. There are four possible combinations of the 2-bit format: (1,1), (1,0), (0,1), and (0,0) which represent the destination addresses destined to both groups, even group, empty, and odd group.

The addresses at level 2 consist of 4 bits which are used for routing at level 2. The first 2-bit field is associated with the even group at level 1 and the second 2-bit field is associated with the odd group at level 1. Examine the first 2-bit field. The subeven group within the even group at level 1 consists of the addresses whose module 4 results are 0 and the subodd group within the even group at level 1 consists of the addresses whose module 4 results are 2. Examine the second 2-bit field. The subeven group within the odd group at level 1 consists of the addresses whose module 4 results are 1 and the subodd group within the odd group at level 1 consists of the addresses whose module 4 results are 3.

In general, for a switching network with size N, the addresses at level m consist of 2^m bits, where $1 \leq m \leq \mathrm{Log}_2 N$. The size of the multicast routing tag is 2N - 2.

It can be observed that at stage 1 of the multicast banyan network there is no blocking if only one of the following three situations is allowed to occur at each switching element.

- one packet which destined to both groups and the other packet is an empty packet.
- two packets where one packet is destined to one group and the other is destined to the other group



	Level 3								Level 2				Level 1	
Routing Tag Format:														
Example: (0, 4, 6, 1)	X	X	1	0	0	1	1	1	1	0	1	1	1	1
(7)	0	0	X	X	X	X	X	X	0	0	X	X	0	0

X: don't care

Figure 20. Tree Hierarchy of the Multicast Routing Field

- one packet which destined to only one group and the other packet is an empty packet.

In order to achieve the above objective, a sorting network is used to rearrange the pattern of the arriving packets. The sorting network sorts the packets using the 2-bit field at level 1. Let the sorting network sort the packets into non-ascending order. After the sorting procedure, the sequence of the packets appears at the outputs of the sorting network is: both groups, even group, empty, and odd group.

Using a shuffle interconnection to connect from the outputs of the sorting networks to the inputs of stage 1 of the banyan network, it is guaranteed that there is no blocking at stage 1 (see Figure 21).

It has been shown that there is no blocking at level 1 of the network, where level 1 consists of one sorting network with size N and stage 1 of the banyan network.

The operation of each switching element at stage 1 of the banyan network is described as follows. The switching element routes the packet to the upper link if the 2- bit tag is destined for the even group; it routes the packet to the lower link if the 2-bit tag is destined for the odd group; it routes and copies the packet to both links if the 2-bit tag is destined for two groups. The empty packet is deleted if the other packet at the other input is destined to both groups; otherwise, the empty packet is sent to the next level. In summary, the 2-bit routing bits at level 1 are used for sorting for the N x N sorting network and routing for stage 1 of the banyan network.

After level 1, the packets have been divided into two groups according to the destination routing tags; the packets destined to the even group are routed to the upper subnetwork and the packets destined to the odd group are routed to the lower subnetwork. Level 2 of the routing tag is used for routing at level 2 of the network which consists of two sorting networks with size N/2 in parallel and stage 2 of the banyan network. The upper subnetwork (or the lower subnetwork) consists of one sorting network with size N/2 and the upper half (or the lower half) of stage 2 of the banyan network.

The upper subnetwork with size N/2 uses the first 2 bits at level 2 of the routing tag, and the lower subnetwork with size N/2 uses the second 2 bits at level 2 of the routing tag for routing. The same routing procedure as in level 1 is applied at each subnetwork.

This operation is repeated at every level until the last level. At the last level, the size of each subnetwork is 2. Hence, no sorting network is required in this level. The last level of the network only consists of stage Log_2 N of the banyan network.

The output ports of the switch check the routing tag of the arriving packet to determine if it is an empty packet or not. If it is an empty packet, it will be discarded. The logic to perform this operation is very simple, which only needs to check a 2-bit field.

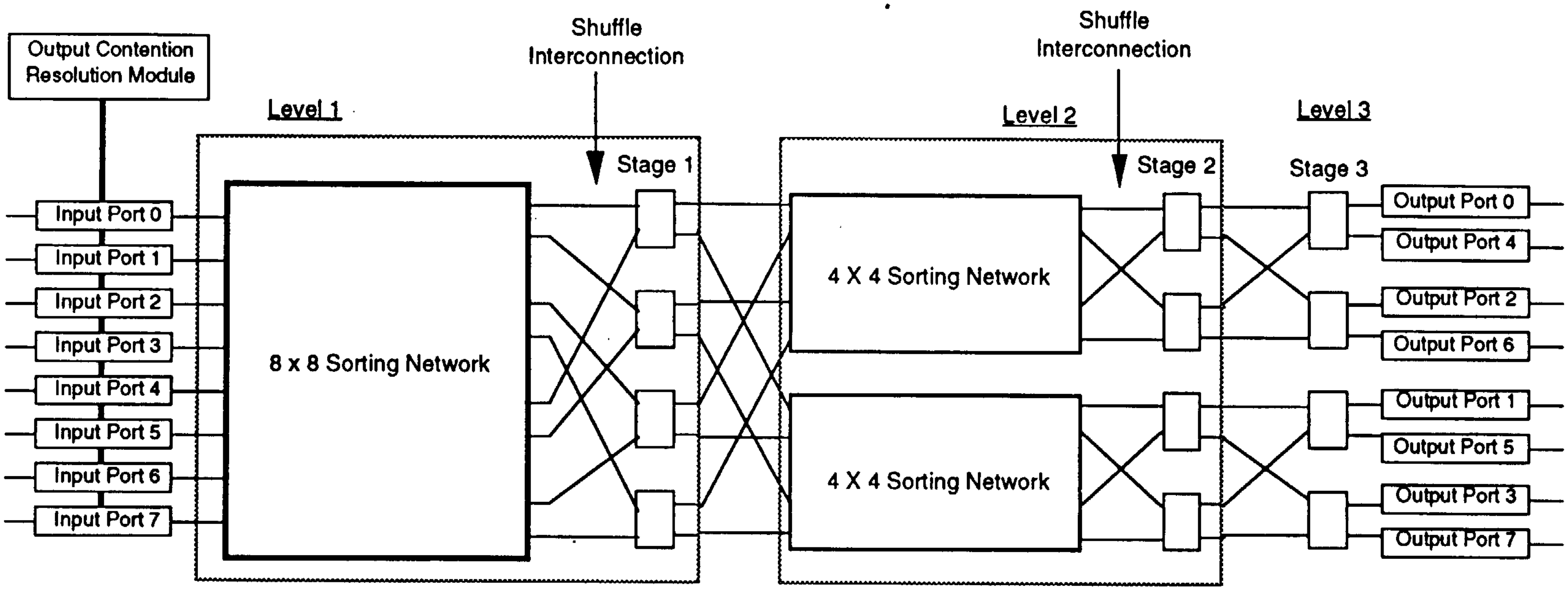


Figure 21. 8 x 8 Nonblocking Multicast Banyan Network

Congestion Issue

Increasing switch speed is often necessary to improve the throughput of the switch. In this case, accumulation of packets may occur at the input port or at the output port. The shift of congestion between input port and output port may be an effective scheme for point-to-point connections; however, for point-to-multipoint connections, this scheme may not be effective or become very complicated. For example, if one of the destined output ports of a multicast packet is in congestion, shall we delay the transmission of the multicast packet to the congested output port only or shall we delay the transmission of the multicast packet to all the destined output ports. This is not a simple problem since the quality of service for all the point-to-multipoint connections may have to be satisfied concurrently. It is suggested that congestion control of a multicast switch be examined in detail in the future.

c. *Multicast Modules at the Output Port*

In this approach, there are multiple multicast modules at the output ports. All the multicast packets are relayed to these multicast modules first through a point-to-point nonblocking switching fabric. And then the multicast modules send the multicast packet to the destined output ports through a point-to-multipoint nonblocking switching fabric (see Figure 22). The number of multicast modules required depends.on the amount of multicast traffic.

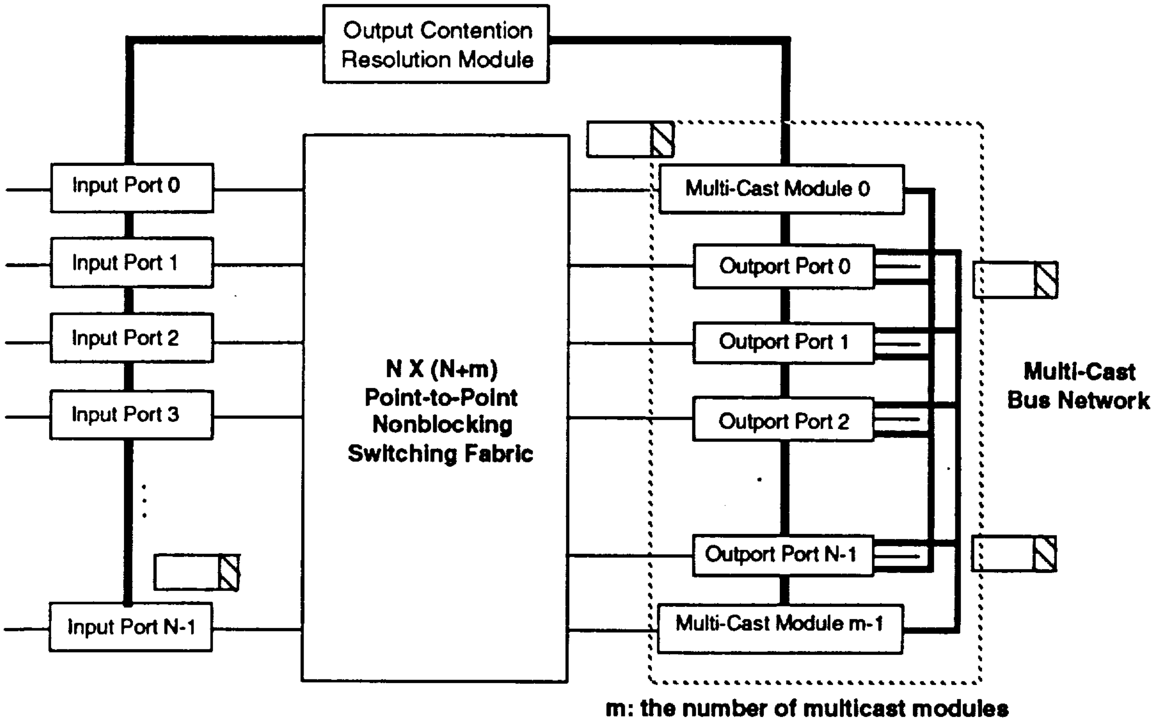


Figure 22. Multicast Modules at the Ouput Ports

The multicast knockout switch uses a similar approach [16] as shown in Figure 23. The knockout switch uses the bus approach to interconnect the inputs and outputs. There are N broadcast buses in the switch for the point-to-point applications. For point-to-multipoint applications, extra multicast modules are required. If there are M multicast modules, then the total number of buses is N + M and the size of the switch becomes N x (N+M). There are (N+M) filters at each bus interface of the output port, where each filter is for one input; hence, the total number of filters for the switch are $N^2 + NM$. It can be seen that the complexity of the bus interface is very high. The desired point-to-point switching fabric is the banyan-type network, which is assumed to be the switching fabric in the discussion below. If the banyan-type network is used as the switching fabric, then the number of filters necessary for the bus interface at each output port is only M, where M is the number of multicast modules.

The output port reservation scheme such as the ring reservation scheme is coupled with the multicast module scheme so that the output port reservation scheme can be done not only for point-to-point connections but also for multicast connections. The multicast module is treated as one of the input ports by the output reservation module. The start of the token stream alternates among N input ports and m multicast modules. Some multicast destination ports are free and some are busy during the output reservation process. As before, it is assumed that the multicast module has the call splitting capability such that the transfer of the multicast packet can be partially completed. In this case, a multicast packet may have to use several slots to complete the transmission of the packet to different destinations.

Congestion Issue

Depending on the traffic distribution, accumulation of packets may occur at the input port or at the multicast module, but not at the output port. If the switch speed is increased, accumulation of packets may also occur at the output port. Since there are three modules involved in congestion control, multicast module at the output port scheme may have the most complicated congestion control procedure.

3.2.2 Path Setup Scheme

For a blocking switching fabric, the packet transfer scheduling algorithm has to be able to resolve output contention and internal blocking at the same time. There are two basic schemes of implementing the packet transfer protocol:

- setup packet phase + transfer packet phase protocol with increased switch speed (or multiple parallel switching fabrics) using input buffering/output buffering.

- transfer packet phase + retransmission packet phase with increased switch speed (or multiple parallel switching fabrics) using input buffering/output buffering.

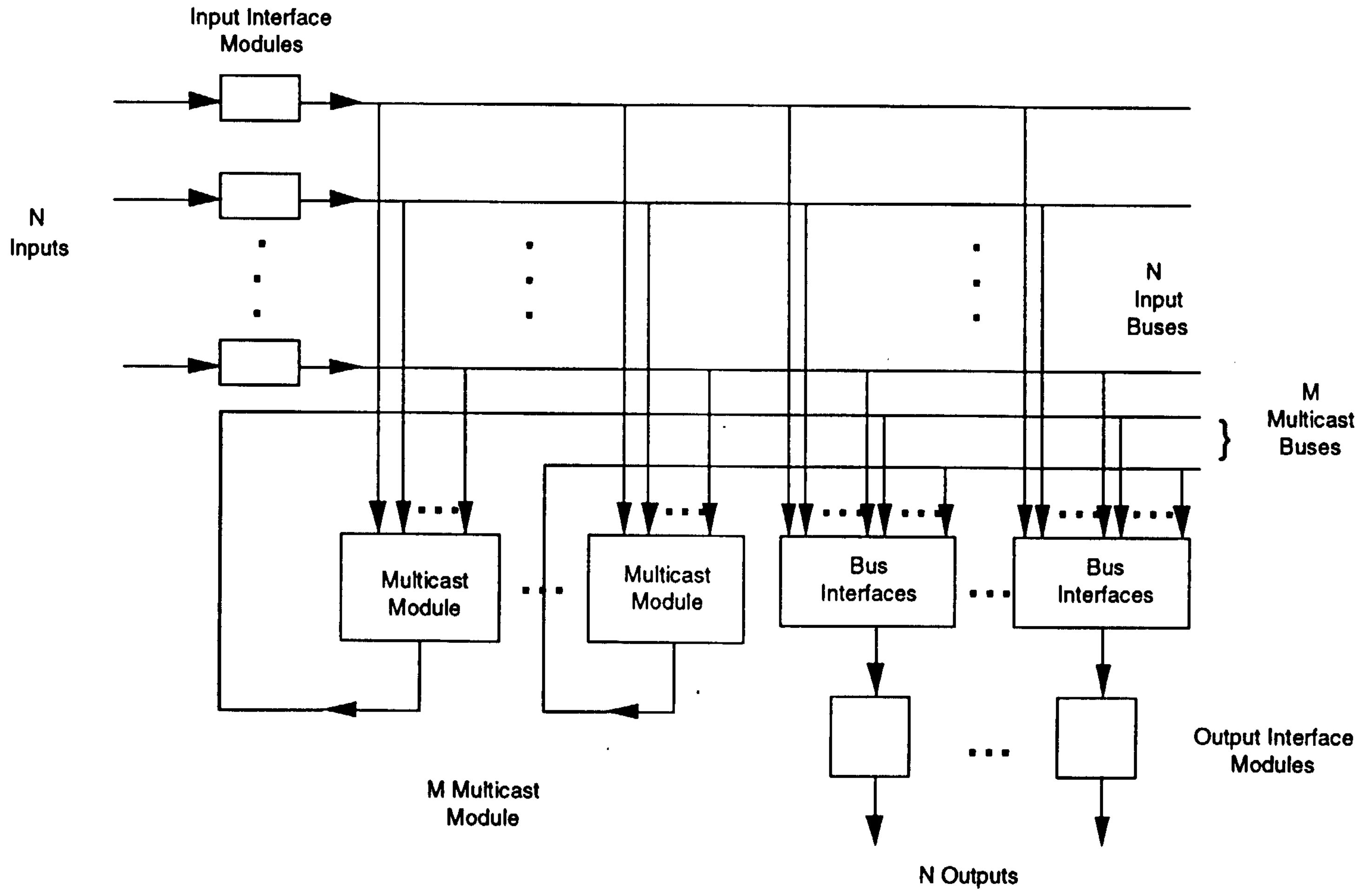


Figure 23. Multicast Knockout Switch

A brief description of the above protocols has been provided in the previous subsection. Both protocols are very similar. The setup packet phase + transfer packet phase protocol is used as the representative to describe the operation of the switches in this subclass.

For a blocking switching fabric, input queueing is a necessity to perform the packet transfer protocol. Several examples are described below.

3.2.2.1 Unbuffered Banyan Switch with an Increased Switch Speed

The operation is described in the following section for a more general case. The maximum throughput of a switch with a blocking banyan switching fabric for different sizes has been reported in [6]. According to [6], the throughput of an 8 x 8 switch is about 0.51; the throughput of a 16 x 16 switch is about 0.45; the throughput of a 32 x 32 switch is about 0.40; the throughput of a 64 x 64 switch is about 0.36. These throughputs are too low to have any practical applications. One way of improving the throughput is to operate the switch at a higher speed. If the switch is operated n times faster than the link speed, then each packet at the input port has n chances to try to set up a path through the switch within one link slot time. Hence, the throughput of the switch is greatly increased. Since more than one packet can arrive at one output port within one link slot time, output queueing is necessary to hold the packets.

3.2.2.2 Parallel Unbuffered Banyan Switches

In this scheme there are *p* copies of banyan networks stacked in parallel and there are *p* transmitters at the input port, *p* receivers at the output port, and output buffering. This switch can be operated in two ways. The first approach is introduced below. In the path setup phase, the setup packets at the input port are loaded into different transmitters. To avoid out-of-sequence transmission, only the packets with distinct destination addresses can be loaded into the transmitters. The setup packets at different transmitters are sent to different copies randomly at the same time. Since the output port has multiple receivers that can receive more than one packets from different input ports at the same time, the throughput is increased and output buffering is required.

In the second aproach, there are *p* minislots reserved for the packet setup phase. These *p* minislots are considered system overhead. For each input port, at minislot 1, the packet at the first transmitter tries to set up a path using the first copy. If the packet encounters blocking either at the switching fabric or at the output port, the packet uses the second copy to set up a path at the second minislot. If the packet successfully sets up a path at minislot 1, then the packet at the second transmitter can use the second copy to setup a path at minislot 2; and so on. In this sequential searching algorithm, the maximum number of reserved minislots to setup a path for the packet at the first transmitter is *p*. The maximum number of reserved minislots to setup a path for the packet at the second transmitter is *p*-1; and so on. Note that corresponding to each minislot of the setup phase, a different copy is used for setting up the path for a packet.

3.2.2.3 Unbuffered Multicast Banyan Switch

The size of the routing tag of a multicast packet used in a blocking multicast banyan network is N [17], where each bit in the routing tag is associated with each output port. There are two registers, one for each output, holding control bits at each switching element (see Figure 24). The operation of each switching element is to AND the routing tag of the packet and the control bits at each register. If the result of the AND operation is 1, a copy of the packet is sent to the output. If the results of the AND operation for both registers are all 1, a duplication of the packet has been made in the switching element

Although the packet transfer protocol of a point-to-point blocking banyan switch can also be applied to the multicast blocking banyan switch, the packet transfer protocol of the latter is more complicated than that of the former. The major differences are as follows. The multicast setup packet has to carry a multiple-destination routing tag. As mentioned above, the size of the routing tag is N. For the point-to-point blocking switch, the ACK signal sent from the output port to the originating input only needs one bit. However for the point-to-multipoint blocking switch, the ACK signals will be sent back from more than one output ports to the originating input port. Since there are more than one ACK coming back to the originating input port, if these ACKs are all sent back at the same time, they will either collide at some switching element or at the input port and lose the information contained in the ACK signal. One way of avoiding conflict at the switching fabric and at the input port is that the TDMA scheme is applied for the ACKs from different output ports to the original input port. The packet setup phase consists of two parts. The first part is that every input port sends an N-bit routing tag through the switching fabric. The second part is that every output port sends back the ACK to the originating input port using the assigned minislot. Output port 0 sends its ACK back to the originating input port using minislot 0; output port 1 sends its ACK back to the originating input port using minislot 1; and so on. Correspondingly, the input port checks whether there is an ACK at each minislot time and determines the packet transfer sequence of the multicast packet for the next slot time.

Congestion Issue

Congestion control of the blocking switch is very similar to the nonblocking switch. When congestion occurs, there are two ways of relieving congestion. The first one is to shift congestion to an uncongested port. The second one is to throttle the sending earth station by continuously broadcasting the queueing length information to all the ground stations. However, the congestion control procedure is complicated by the multicast operation. As discussed before, the key issue is what do we do about the multicast packet if one of the destined output ports is in congestion.

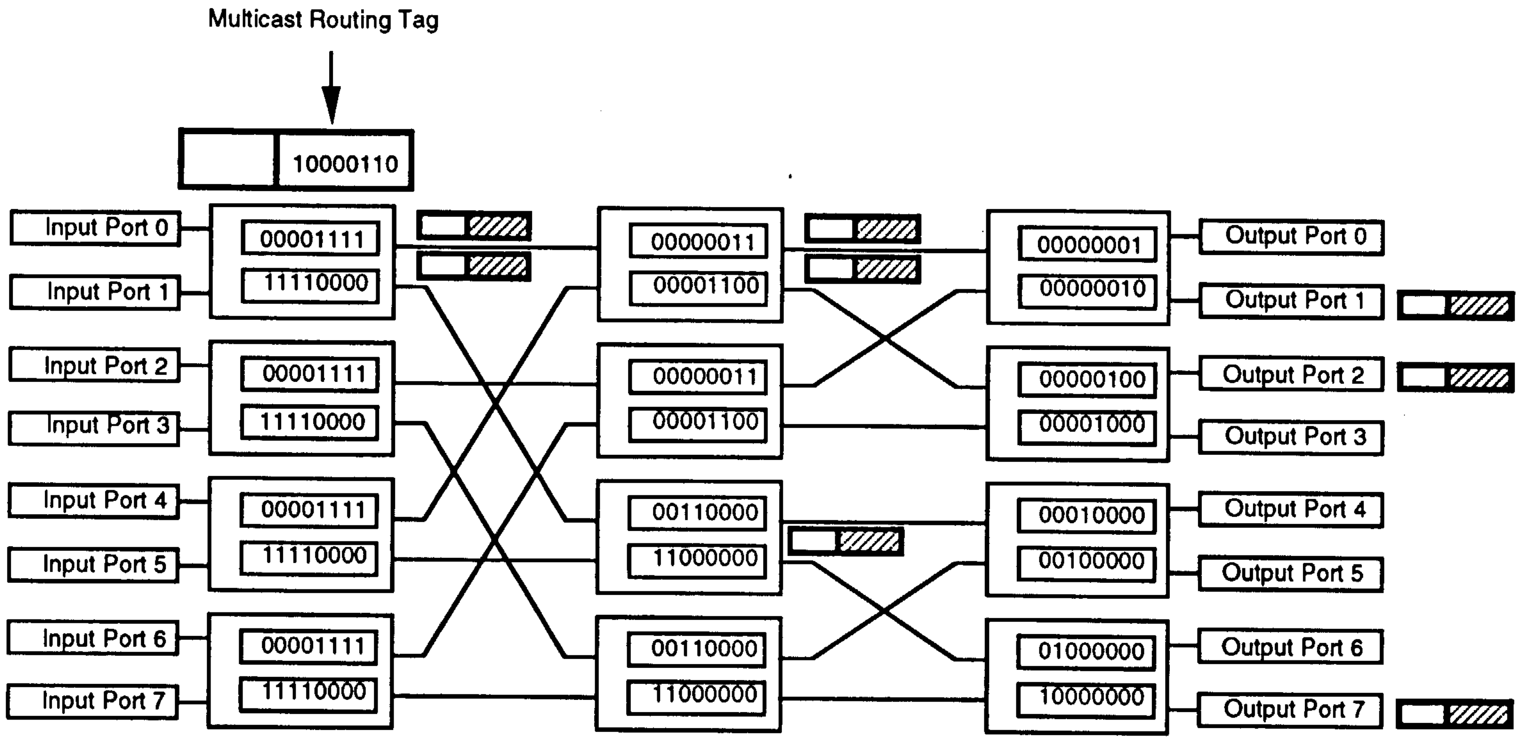


Figure 24. Multicast Banyan Switch

3.2.3 Address Filter Scheme

In this subclass of switches, there is a disjoint path between each input and output pair. Thus, the switching fabric is both point-to-point and point-to-multipoint nonblocking. The packet transfer protocol uses the forward-and-store approach. The forward phase is used to transmit the arriving packets to different output ports since the packets are not stored at the input ports. Each output has filters to select the packets destined to itself. To resolve output port contention, more than one packets are allowed to arrive to one output port at the same time; hence, multiple receivers and output buffering are required to store the packets. Note since the switching fabric is nonblocking, there is no internal blocking problem and since the buffers are located at the output ports, there is no head of line blocking. For the above reasons, the throughput of a switch with output buffering is higher than that with input buffering [5]. Two examples are provided as follows.

3.2.3.1 Knockout Nonblocking Switching Fabric with Output Buffering

The knockout switch shown in Figure 25 uses the bus approach to interconnect the inputs and outputs [16]. There are N broadcast buses, one from each input port, in the switch, and there are N filters at each bus interface of the output port. The total number of filters for the switch is N^2.

Since there is a disjoint path between any input-output pair in this topology, there is no internal blocking. The packet transfer protocol uses the forward-and-store scheme with output buffering; hence, there is no HOL blocking. The N filters at each output port performs as N receivers which can receive N arriving packets at the same time. After the N receivers, there is one output buffer which performs as a statistical multiplexer. The amount of buffering required at each output port depends on the packet loss ratio requirement.

Congestion Issue

Since only output buffering is employed, congestion only occurs at the output port. Congestion control can be achieved by monitoring the output queue length continuously and broadcasting the information to all the ground stations. The ground station delays the transmission of the packet whose destined downlink beam is in congestion.

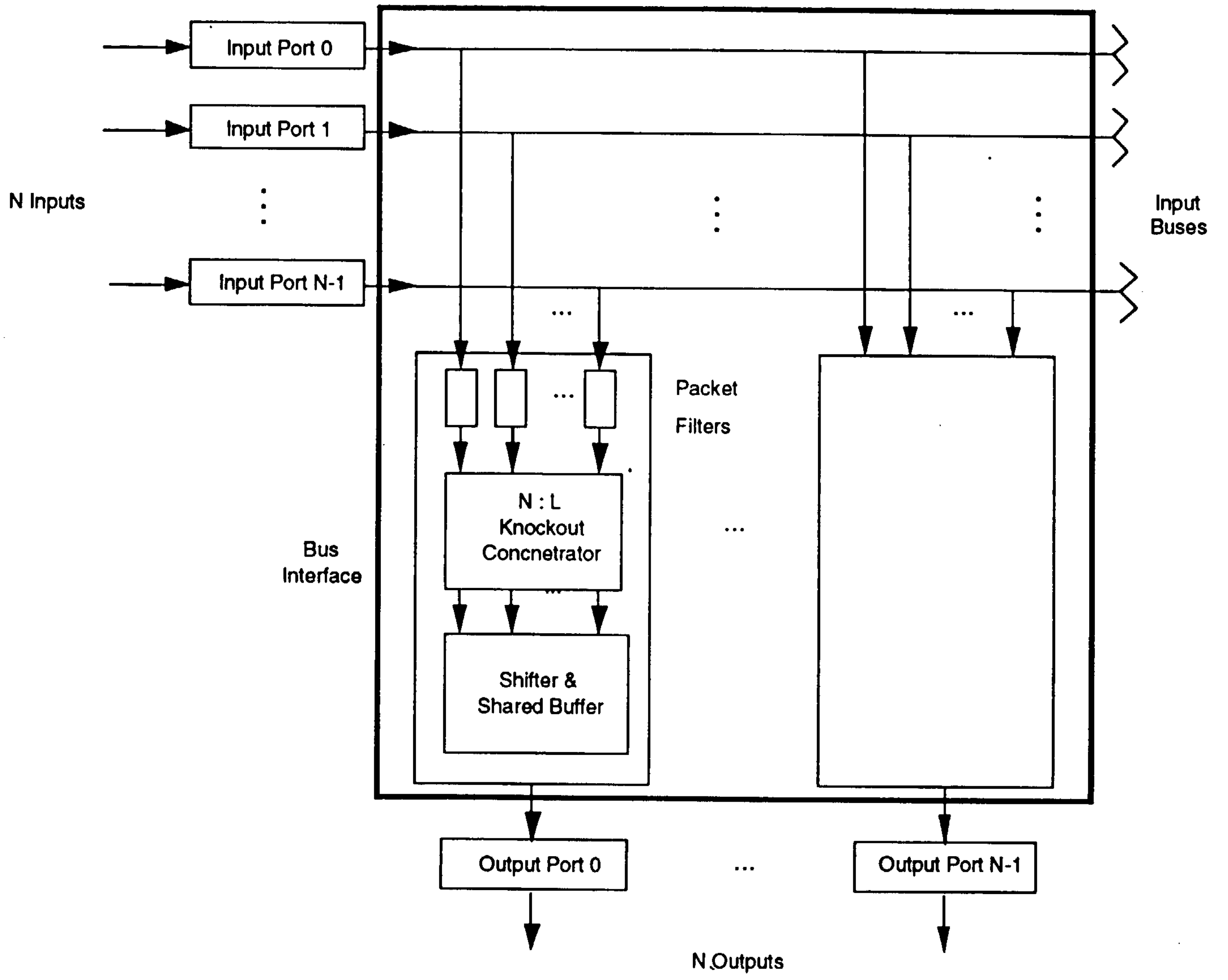


Figure 25. N x N Knockout Switch with Ouput Queueing

3.3 Throughput Performance

This subsection addresses the throughput performance of two switches using computer simulation techniques. These switches are:

a. point-to-point nonblocking switching fabric with input buffering.

b. point-to-multipoint nonblocking switching fabric with input buffering.

The impact on the switch throughput resulting from an increased switch speed and an improved output contention algorithm is analyzed. The effect of traffic imbalance on the throughput is also studied.

3.3.1 Simulation Models

The simulation is based on discrete-event simulation. The simulation is performed on a SUN SPARC workstation using the OPNET simulation package from MIL 3, Inc. Two switch models are described below.

Model A:

- switch size: 8 x 8
- switching fabric: point-to-point nonblocking
- switch buffering: input
- output contention resolution scheme: input ring reservation

Model B:

- switch size: 8 x 8
- switching fabric: point-to-multipoint nonblocking
- switch buffering: input
- output contention resolution scheme: input ring reservation

The simulation model shown in Figure 26 consists of traffic generators, input ports, a switch fabric, output ports, and a token generator. The traffic generators generate packets following Possion distribution. The input port stores the arriving packets. The token generator associated with the input ports performs output reservation for the arriving packets. The switch fabric routes the packets to the destined output ports. Depending on whether the switch speed is higher than the input link speed, the function of the output port is different. If the switch speed is equal to the link speed, the output port is only a sink. If the switch speed is higher than the link speed, the output port performs as a statistical multiplexer with a FIFO queue.

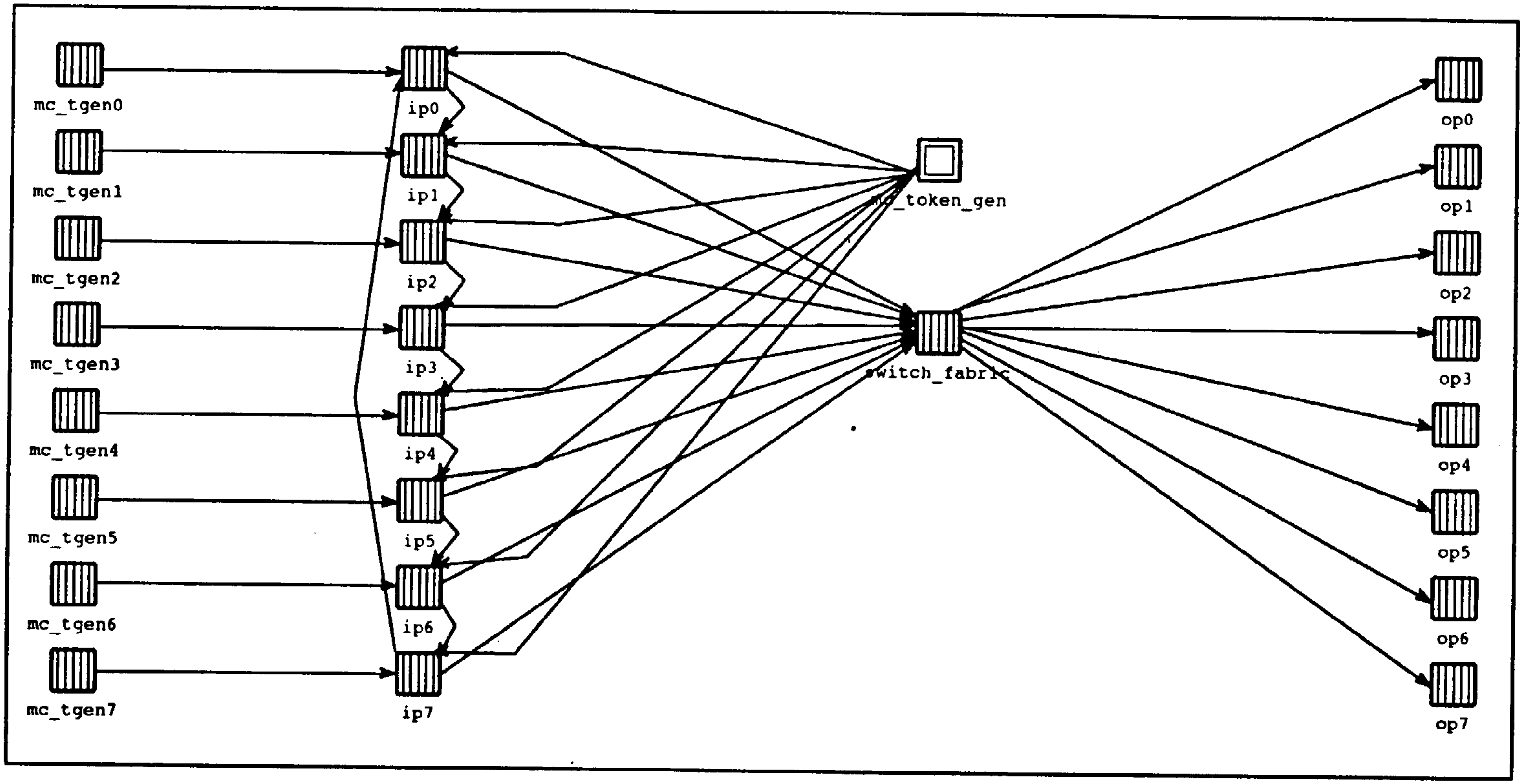


Figure 26. Switch Simulation Model

The throughput of a switch with input queueing is limited due to the head of line blocking problem. As previously mentioned, one method to improve the throughput is to use a non-FIFO queue. If the first packet is blocked due to output blocking, the output reservation algorithm examines the packets at the back of the first packet in the queue. The number of packets examined each time (or the checking depth) is a parameter. The other method to improve the throughput is to increase the switch speed, which is another parameter.

The parameters to be varied for Model A are as follows:

- switch speedup, i.e., switch speed/link speed (s)
- input link utilization
- checking depth (d).

All the discussions about the point-to-point switch simulation model can be applied to the point-to-multipoint switch simulation model, since the former is a special case of the latter. The traffic generator generates two types of packets following a certain distribution. One is the point-to-point packet and the other one is the multicast packet. For a point-to-point switch with a given input link utilization (assume that the input link utilizations at different input ports are uniform), the traffic intensity through the switch is determined, i.e., the average link utilization is the traffic intensity. For a multicast switch, three more parameters are required to determine the traffic intensity. The first one is the multicast packet ratio defined as the ratio of the number multicast packets to the total number of arriving packets. The next two parameters are the lower bound and the upper bound of the number of destinations each multicast packet carries. The lower bound is always larger than or equal to 2. The upper bound is always smaller than or equal to the switch size. For simplicity, the lower bound is assumed to be 2 and the upper bound is assumed to be the switch size, i.e., 8. Assume that the number of destinations each multicast packet carries follows the uniform distribution between the lower bound and the upper bound. Given a multicast packet ratio (*mr*), multicast lower bound (2), multicast upper bound (8), and input link utilization (ρ), the traffic intensity (ρ_i) can be calculated as follows.

$$\rho_i = \rho \bullet mr \bullet \frac{2+8}{2} + \rho\,(\,1 - mr) \;=\; \rho\,(1 + 4mr)$$

Note that the value of ρ_i should be always less than 1.

A table of ρ_i is shown in Table 1 for different ρ and *mr*.

Table 1. Traffic Intensity for Different mr

ρ_i	$mr = 0.05$	$mr = 0.1$	$mr = 0.15$	$mr = 0.2$
$\rho = 0.4$	0.48	0.56	0.64	0.72
$\rho = 0.45$	0.54	0.63	0.72	0.81
$\rho = 0.5$	0.60	0.70	0.80	0.90
$\rho = 0.55$	0.66	0.77	0.88	0.99

Parameters to be varied for Model B are given below:

- switch speedup, i.e., switch speed/link speed (s)
- input link utilization
- checking depth (d)
- multicast packet ratio (mr).

3.3.2 Simulation Results

Simulation has been conducted for an 8 x 8 switch. Although ATM parameters are used in simulation (a switch speed of 155.52 Mbit/s and a packet size of 424 bits), the simulation results presented herein are applicable to other system parameters as well (i.e., the simulation results are not affected by a particular switch speed or packet size).

The objective of the simulation is to obtain the saturation throughput of the switches. Without increased switch speed, the throughput is defined as the average number of packets arriving at the output ports in one link slot divided by the switch size, where a link slot is defined as (packet size/input link speed). The input buffer size is assumed to be infinite.

The first set of results shows the effectiveness of a larger checking depth to improve the throughput. As seen from Figure 27, the improvement of throughput gets less when the checking depth gets larger. For a checking depth of 2, the throughput can reach 0.73; for a checking depth of 3, the throughput is 0.79; for a checking depth of 4, the throughput is 0.83. For a real application, it is not cost effective to use a very large checking depth to improve the throughput. The best approach is to use a small checking depth such as 3 or 4 and to increase the switch speed to make the switch throughput close to 1.

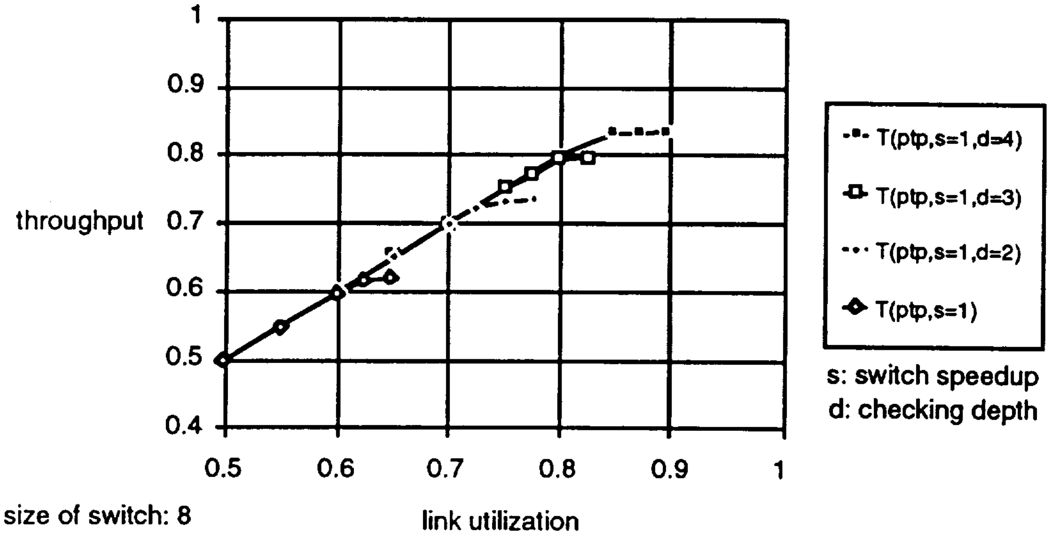


Figure 27. Throughput Performance of the 8 x 8 Point-to-Point Switch for Different Checking Depths

The second set of results shows the effectiveness of increasing switch speed to improve the throughput. As shown in Figure 28, the improvement of throughput is proportional to the increase of the switch speed, which is very effective.

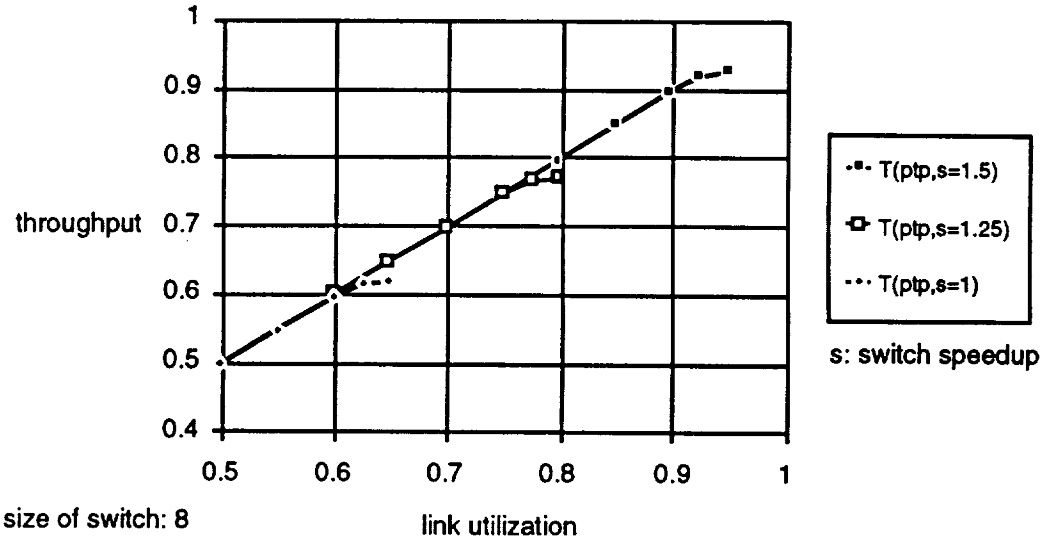


Figure 28. Throughput Performance of the 8 x 8 Point-to-Point Switch for Different Speedups

The third set of results shows the effectiveness of a larger checking depth for the point-to-multipoint switch (see Figure 29). As in the point-to-point switch case, the improvement of throughput gets lesser as the checking depth gets larger. For a checking depth of 4 and a multicast packet ratio of 0.1, the saturation throughput can achieve 0.89.

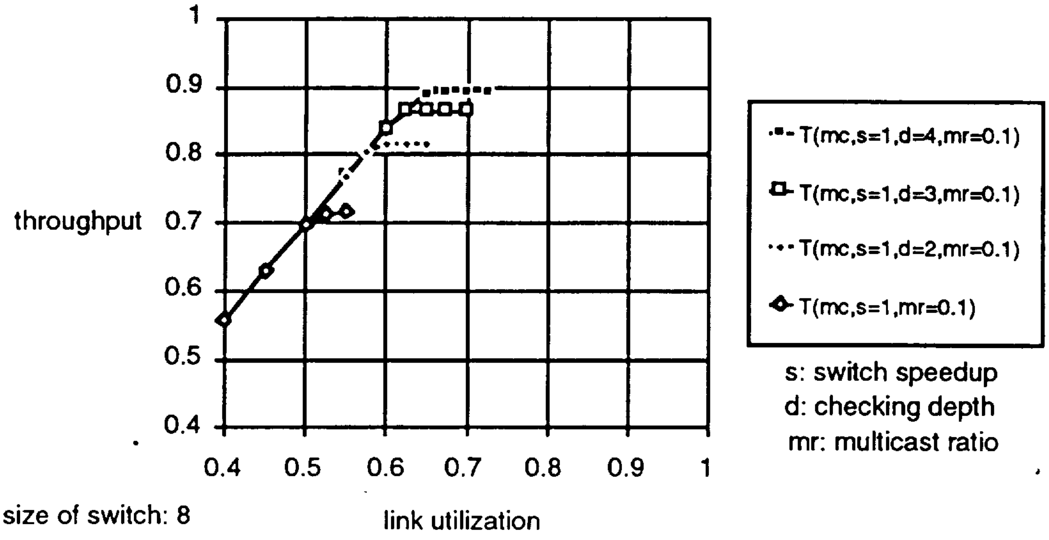


Figure 29. Throughput Performance of the 8 x 8 Point-to-Multipoint Switch for Different Checking Depths when the Multicast Packet Ratio is 0.1

The fourth set of results shows the effectiveness of increasing switch speed for the point-to-multipoint switch. The improvement of throughput is proportional to the increase of the switch speed, which is very effective. Figure 30 shows the throughput performance when the multicast ratio is 0.1 for every input port, and Figure 31 shows the throughput performance when the multicast ratio is 0.2. for every input port.

There are two forms of traffic imbalance for the point-to-point switch. The first one is the nonuniform output destination distribution experienced at each input port. The second one is the nonuniform input link utilization among different input ports.

The purpose of this set of results is to illustrate that certain traffic imbalance situations will reduce the saturation throughput, and as a result congestion may occur. It is not intended to enumerate all the possible traffic imbalance situations.

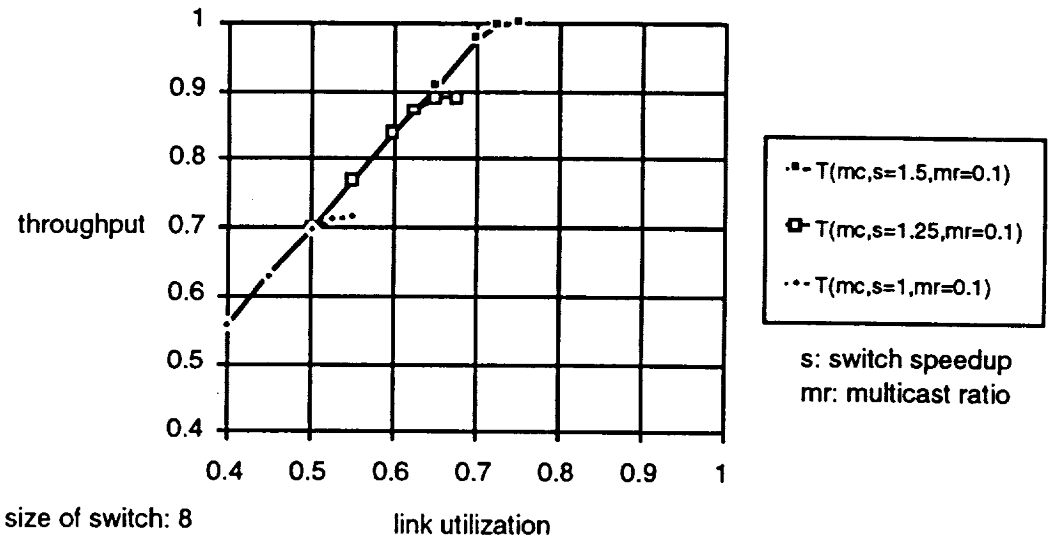


Figure 30. Throughput Performance of the 8 x 8 Point-to-Multipoint Switch for Different Speedups when the Multicast Packet Ratio is 0.1

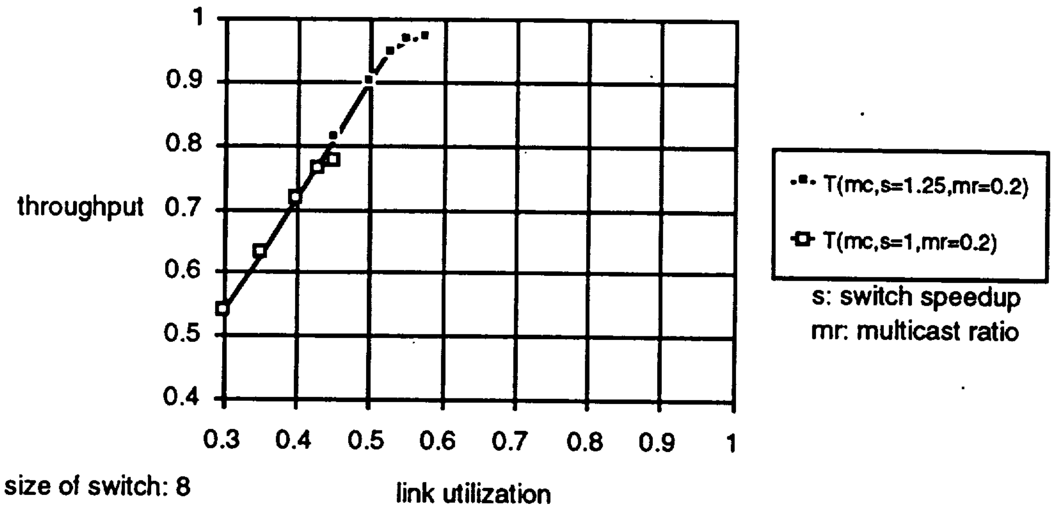


Figure 31. Throughput Performance of the 8 x 8 Point-to-Multipoint Switch for Different Speedups when the Multicast Packet Ratio is 0.2

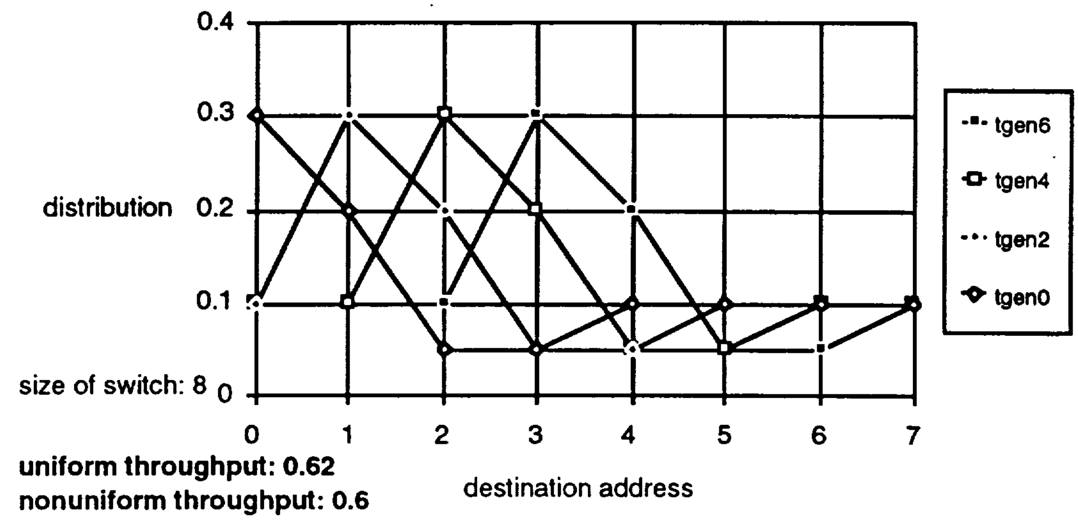
The traffic imbalance situations and their associated saturation throughputs are shown in Figures 32 to 35. Figures 32 and 33 show the effect of nonuniform output destination distributions on the saturation throughput. Although the total incoming traffic intensity is the same as that in the uniform case, the throughput of a switch with a nonuniform output destination distribution is lower than that of a switch with a uniform output destination distribution. It can be seen that if the distribution curve is narrower, the reduction of the throughput is also larger. Figures 34 and 35 show the effect of nonuniform input link utilizations on the throughput. If the mean utilization difference among different input links is larger, the reduction of throughput is also larger. The combination of a nonuniform output destination distribution and a nonuniform input link utilization worsens the throughput performance. The reduction of the throughput is not the sum of the reduction from each case. For example, the reduction of the throughput in Figure 33 is 0.08 and the reduction of the throughput in Figure 34 is 0.074; the reduction of the throughput for the combined effect of Figures 33 and 34 is 0.09.

The traffic imbalance for the point-to-multipoint switch is much more complicated than that of a point-to-point switch. The possible traffic imbalance situations include the following:

- nonuniform output destination distribution
- nonuniform input link utilization
- nonuniform multicast packet ratio
- nonuniform distribution for the number of destinations that each multicast packet carries.

From the simulation, it is found that the saturation throughput is quite insensitive to the multicast packet ratio distribution for different traffic generators. For example, in Figure 36, the reduction of throughput is only 0.01 compared with that of a switch with a uniform multicast packet ratio of 0.1. The effect of the nonuniform distribution for the number of destinations that each multicast packet carries depends on the average number of destinations. For the uniform case, the average number of destinations that each multicast packet carries is $\frac{2+8}{2} = 5$. If the average number of destinations each multicast packet carries for the nonuniform case is larger than 5, the saturation throughput is increased; otherwise, it is decreased.

Note that the purpose of above discussion is to illustrate the effect of traffic imbalance on the saturation throughput for an 8 x 8 switch. To understand the effect of traffic imbalance, traffic correlation, and time varying traffic on the saturation throughput, a further research effort is needed.



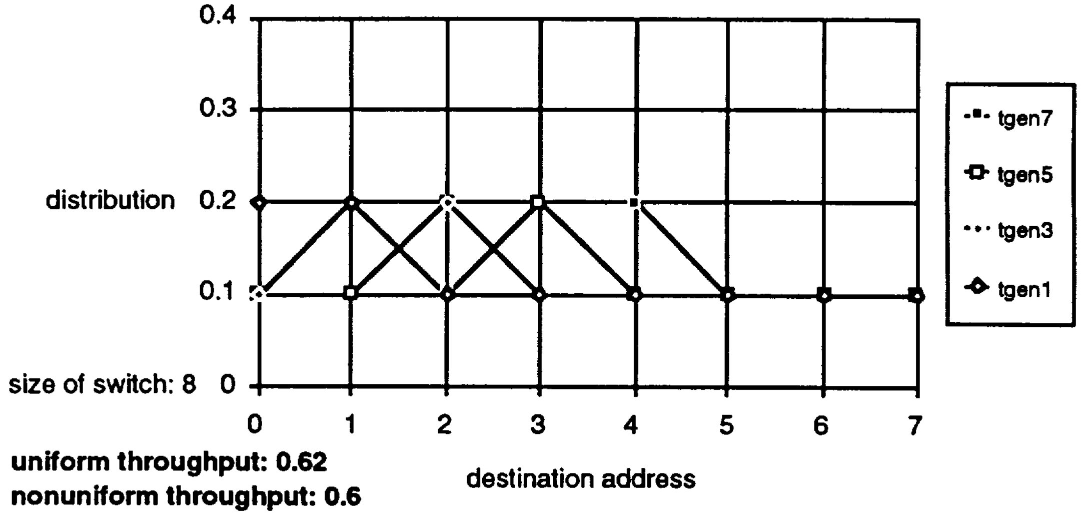


Figure 32. Nonuniform Destination Distributions for 8 Traffic Generators

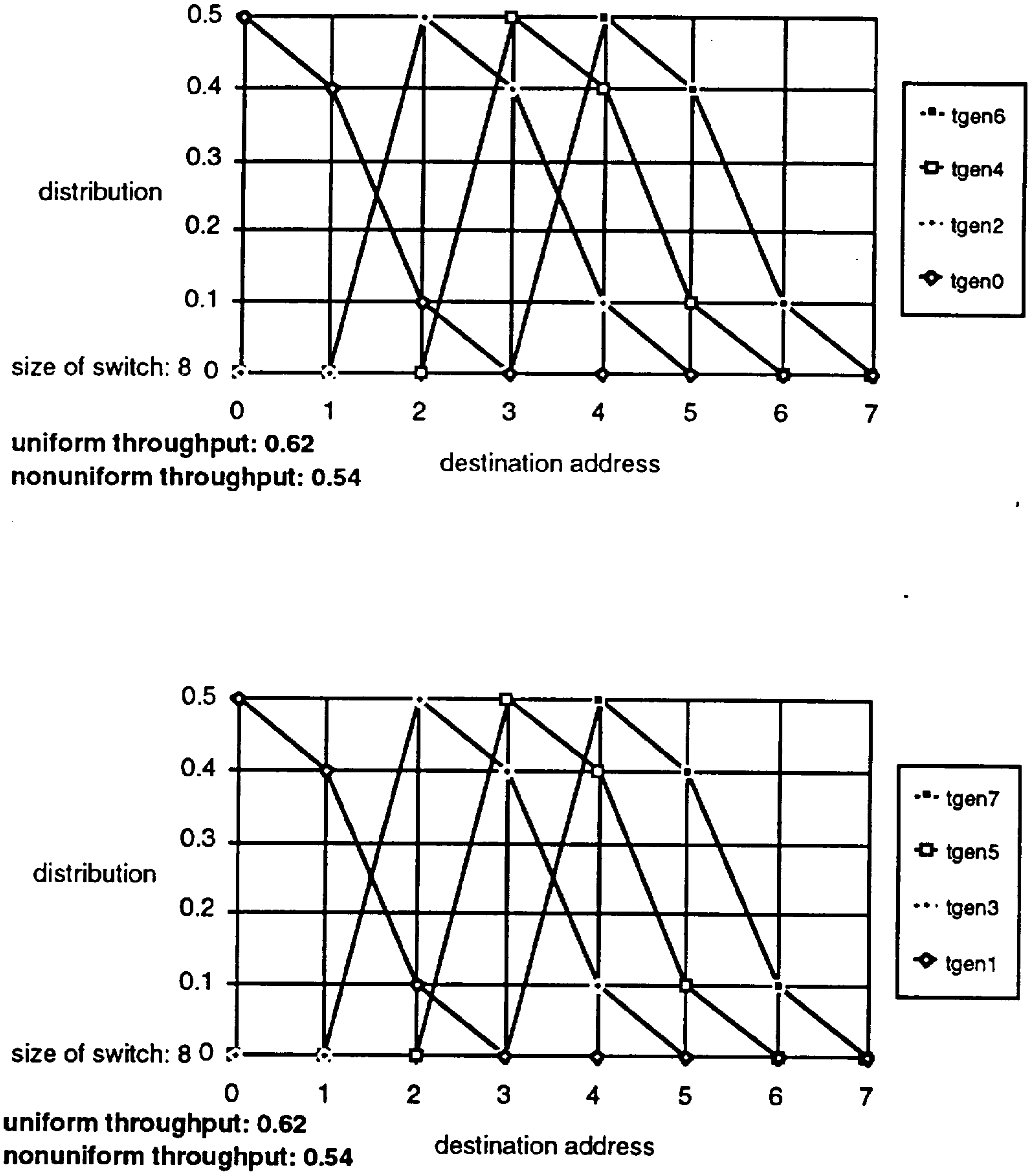


Figure 33. Nonuniform Destination Distributions for 8 Traffic Generators

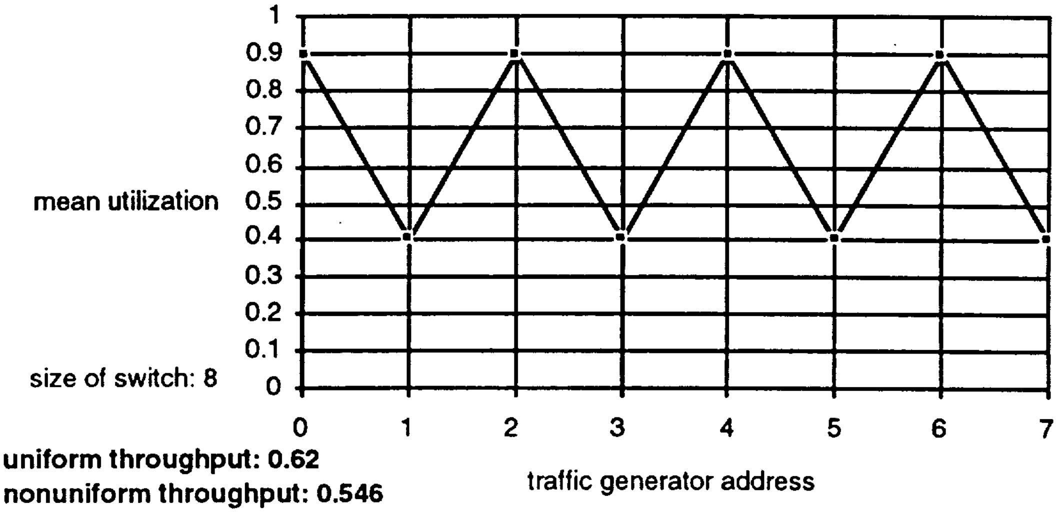


Figure 34. Nonuniform Input Link Utilizations

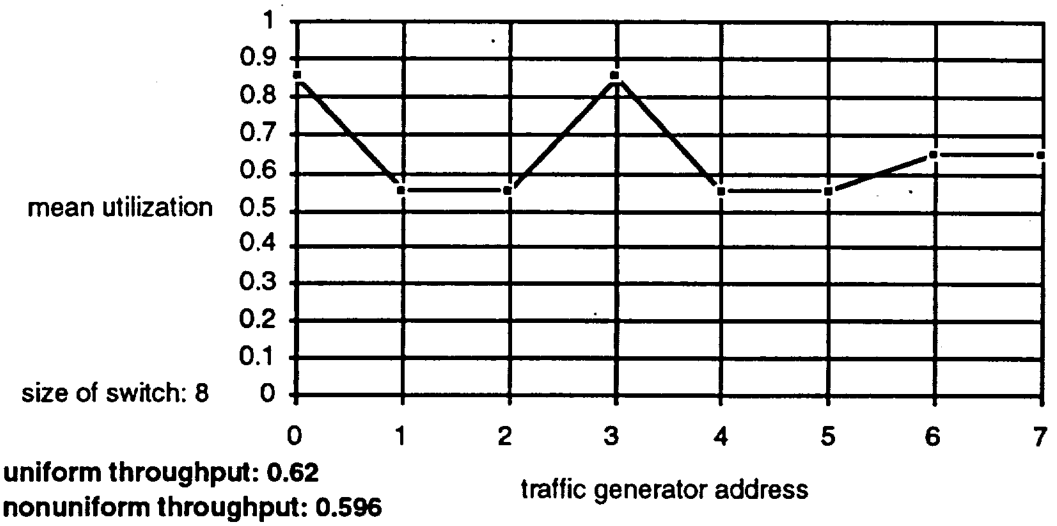


Figure 35. Nonuniform Input Link Utilizations

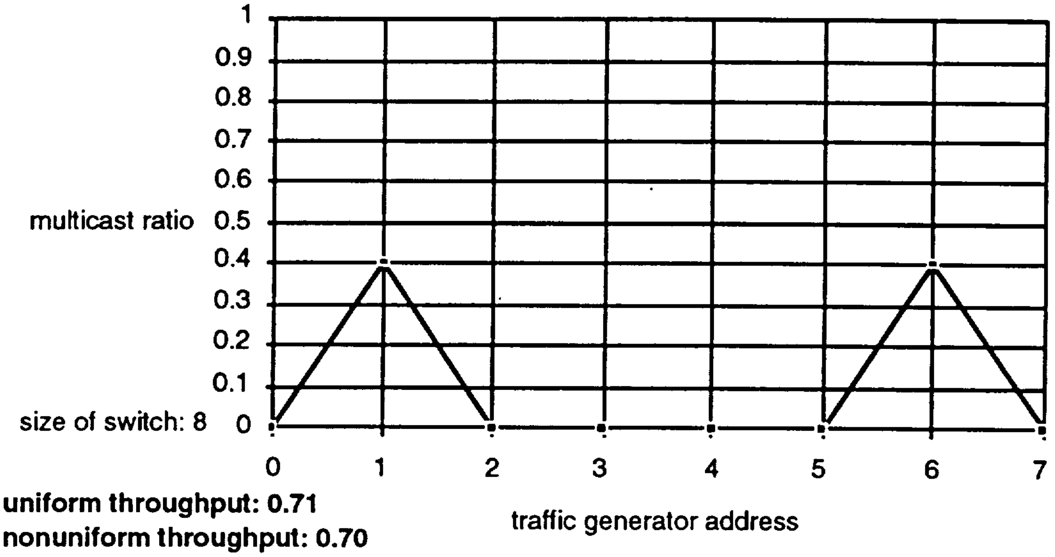


Figure 36. Nonuniform Multicast Packet Ratio for 8 Traffic Generators

3.4 Summary of Switch Contention

Two types of fast packet switching architectures, i.e., contention-free switching architecture and contention-based switching architecture, are considered in the previous sections. The contention-free switching architectures, by definition, are free from contention. This type of switching architecture, however, has a capacity limited to several Gbit/s. This is more than sufficient for the total system capacity required for the 64-kbit/s packet service (590 Mbit/s). However, for a larger system with a capacity of 10 Gbit/s or higher, the contention-based switching architecture is more appropriate. In general, the most common contention-based switching architecture discussed in the literature and implemented in the industry is the multistage switching architecture.

For the multistage switching architectures, scheduling of packet transfer at the input ports is necessary to avoid output contention. Several packet transfer scheduling algorithms are described for both nonblocking and blocking switching fabrics. Among them, the input ring reservation scheme for the nonblocking switching fabric attracts most attention due to its easy implementation and versatile applications. For the multistage switching architecture, the throughput can not reach 1 due to head of line blocking at the input ports. Two schemes to improve the throughput of the multistage switching architecture are discussed. The first one uses a larger checking depth for each input port, and the other is to increase a switch speed. To fully understand the effectiveness of these schemes, simulation is performed.

An 8 x 8 fast packet switch with a nonblocking switching fabric is used as the switch model for throughput performance analysis. Simulation models are built and experimental sets are run to collect the throughput results. From the simulation results, the improvement of throughput is proportional to the increase of the switch speed. For an 8 x 8 point-to-point fast packet switch, the switch speed has to be increased by 65% to reach a throughput of 1. Simulation results also show that a larger checking depth is an effective way of improving the throughput. However, since the improvement of throughput gets less when the checking depth gets larger, it is not practical to use an extremely large checking depth. The best scheme to improve the throughput is to use a checking depth of 3 or 4, and also to increase the switch speed. For an 8 x 8 point-to-point fast packet switch, with a checking depth of 3 or 4, the switch speed has to be increased by 20% to 27% to achieve a throughput of 1. For a point-to-multipoint fast packet switch, the throughput is determined not only by the input link utilization but also by the multicast packet ratio. For an 8 x 8 point-to-multipoint fast packet switch with a multicast packet ratio of 0.1 and a checking depth of 3 or 4, the switch speed has to be increased by 12% to 15% to achieve a throughput of 1.

4 On-Board Switch Output Multiplexing

Traffic channels routed through circuit and packet switches are multiplexed by a TDM frame formatter to generate dwell traffic bursts for each downlink beam. This section addresses a number of system issues involving a TDM frame formatter and alternate design approaches. Based on the discussions in the previous sections, the following assumptions have been made to analyze system design issues:

a. Routing of a circuit switched traffic channel is deterministic for the duration of a call connection

b. Switch contention resolution and address filtering for data packets have been properly performed such that the input to the TDM formatter consists of only the data packets destined to the designated downlink beam

c. Routing of multicast traffic channels (both circuit and packet switched traffic) to different downlink beams has been properly performed such that the TDM formatter only needs to perform multicast traffic processing for the dwell areas within one downlink beam

d. According to Section 2, the system provides the following traffic capacity (Table 2):

Table 2. System Capacity

TRAFFIC TYPE	TOTAL BEAM CAPACITY (Mbit/s)	TOTAL SYSTEM CAPACITY (Mbit/s)
Circuit Switched Traffic	81.92	737.28
Packet Switched Traffic	65.536	589.824
Total Traffic	147.456	1327.104

The capacity values shown in the table do not include signaling channels and control/status messages that are necessary for system operation, and hence actual values may be slightly higher than those given in the table.

e. A baseband switch architecture utilizes output buffering, such as a TDM bus with distributed output memories, a fiber optic ring switch, or a multistage banyan-based switch. (The general discussion presented below is also applicable to other types of switches, such as a common memory switch, but requires some modification.)

The system design issues addressed in the following sections include a downlink TDM frame structure, buffer implementation options, multicast traffic processing, and buffer sizes.

4.1 Downlink TDM Frame Structures

Downlink hopping beam transmission requires a minimum of eight TDM bursts, one for each dwell area. Two types of TDM frame structures are considered in the following.

The first structure, depicted in Figure 37, consists of eight dwell area reference bursts (RBs) and up to eight traffic bursts (TBs). The RB provides a reference timing to all the earth stations within the designated dwell area and includes network control messages for uplink carrier frequency allocation/deallocation, frame numbers, a downlink traffic burst position, circuit slot assignment, and other control/status information. Each RB is assigned a unique identification code to distinguish it from other RBs to the adjacent dwell areas or to the adjacent beams. The TB carriers traffic channels to a designated dwell area. If some dwell area has no traffic, there will be no traffic burst assigned to the dwell area.

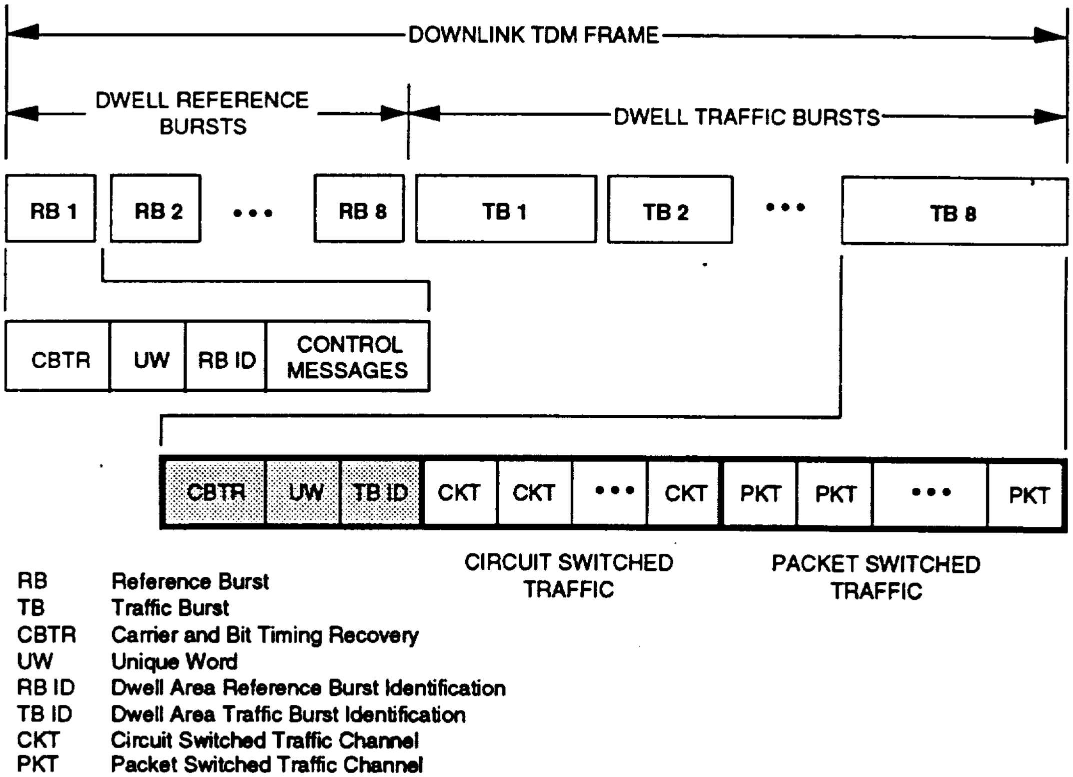


Figure 37. Downlink TDM Frame Structure with Dedicated Reference Bursts

In this frame structure, the RB locations are prefixed and will not be affected by downlink time plan changes. An obvious shortcoming is a less efficient frame utilization due to additional guard times and preambles required for multiple bursts per dwell area.

The second TDM frame structure, shown in Figure 38, overcomes this shortcoming by combining the two types of bursts. There will be exactly eight dwell area bursts in one frame. If there is no traffic to a certain dwell area, only the preamble and control message field will be transmitted to the area.

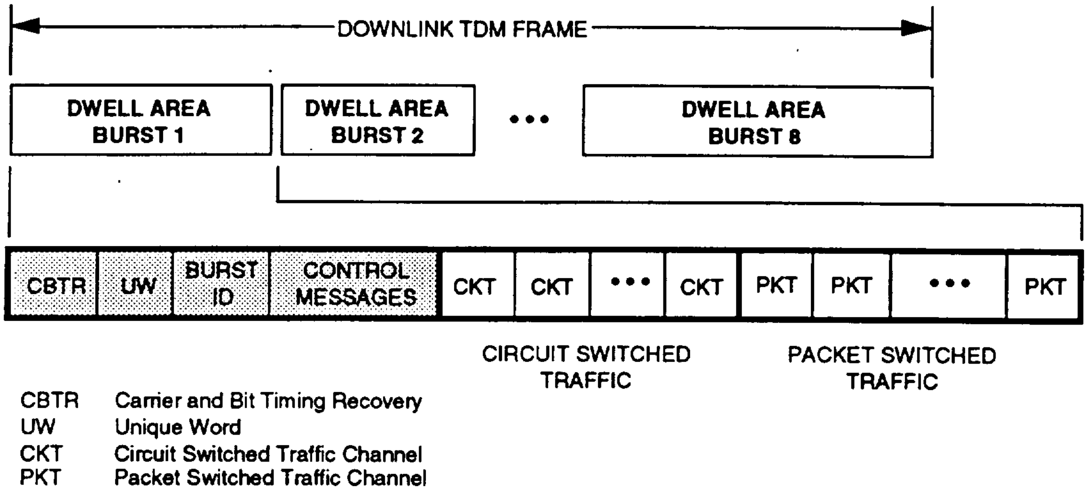


Figure 38. Downlink TDM Frame Structure for Single Burst per Dwell Area

The shortcoming of this frame structure is that the burst positions may change as a traffic volume to one area increases or decreases, and implementation of frequent downlink time plan changes may not be as reliable as in the previous frame structure. In addition, the earth station requires special coordination with the on-board processor (or the network control center) during initial receive timing acquisition such that the given downlink burst does not change its position until the completion of the acquisition process.

Between the two frame structures described above, the first structure is operationally more flexible than the second. To assess the impact of a higher overhead on frame efficiency, consider a burst overhead of 128 bits (guard time, a carrier-and-bit-timing-recovery pattern, and a unique word) and an RB control field size of 128 bits. Table 3 shows a comparison of frame inefficiency resulting from burst and frame overheads for the two types of frame structures. For a frame period of 0.5 ms or longer, the resulting frame inefficiency is less than 5 percent. For a frame period of 250 µs, the frame inefficiency figures for the dedicated RB and single dwell burst techniques are respectively 8.3 and 5.6 percent. The downlink transmission rate must be increased accordingly to maintain the nominal transmission capacity. In general, there is no significant difference in frame efficiency between the two types of frame structures for a frame period of 250 µs or longer, and hence the first TDM frame structure is preferred for implementation.

Table 3. Frame Inefficiency

FRAME PERIOD (ms)	DEDICATED REFERENCE BURST	SINGLE BURST PER DWELL AREA
0.25	8.33%	5.56%
0.5	4.17%	2.78%
1.0	2.08%	1.39%
2.0	1.04%	0.69%
4.0	0.52%	0.35%
8.0	0.26%	0.17%

4.2 Multiplexer Implementation

4.2.1 Implementation Options

The TDM frame formatter includes a buffer to perform a speed conversion from the baseband switch speed to the downlink transmission rate, multiplexing of circuit and packet switched traffic, TDM formatting, and queueing for packet switched traffic. A multiplexer can be implemented using two separate buffers for circuit and packet switched traffic or a single shared buffer. These two approaches are illustrated in Figures 39 and 40. Also included in the figures are an RB and preamble generator and a TDM controller. In actual implementation, the content of the RB (e.g., control messages) is generated by the autonomous network controller (ANC) and routed to the TDM frame formatters through a baseband switch, and a preamble pattern is a fixed bit sequence prestored in the designated memory locations.

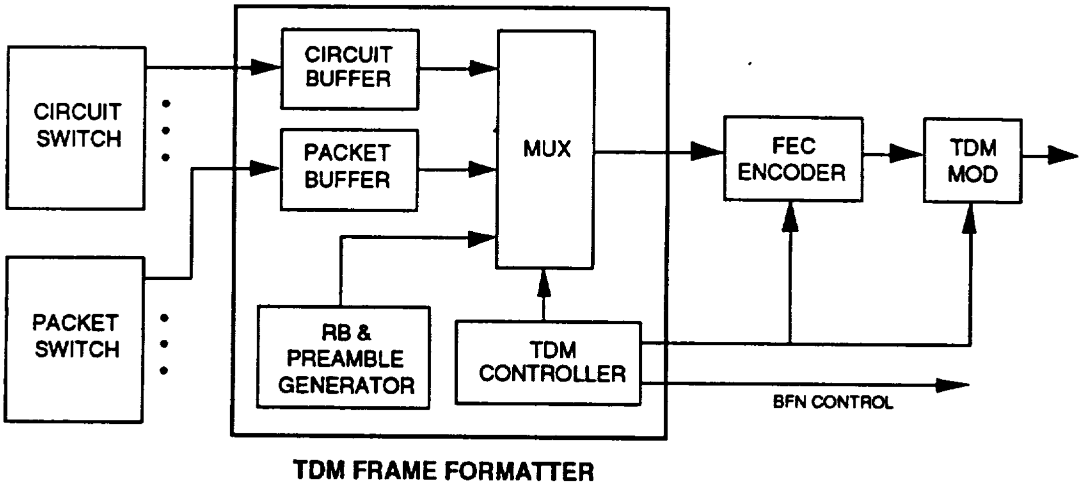


Figure 39. Separate Buffers for Circuit and Packet Switched Traffic

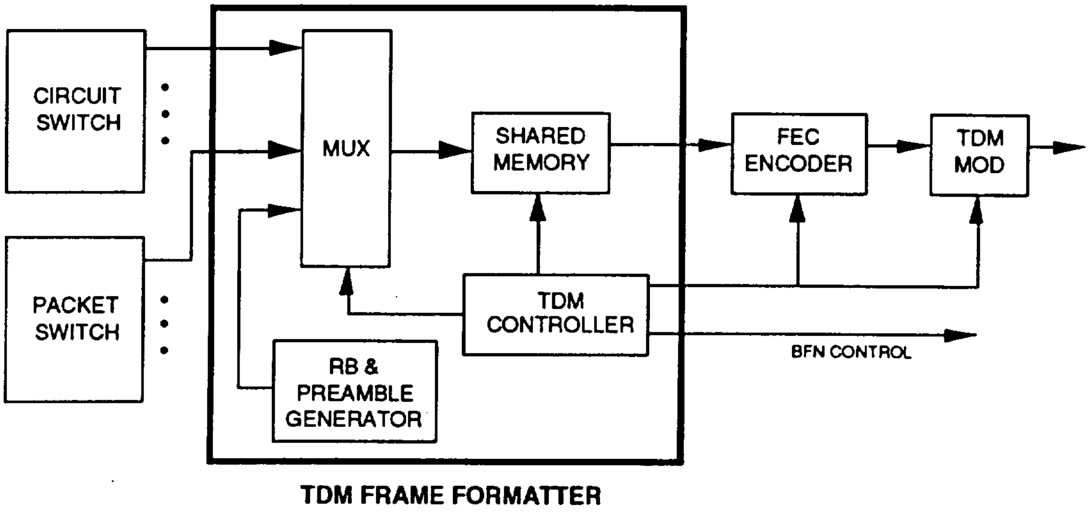


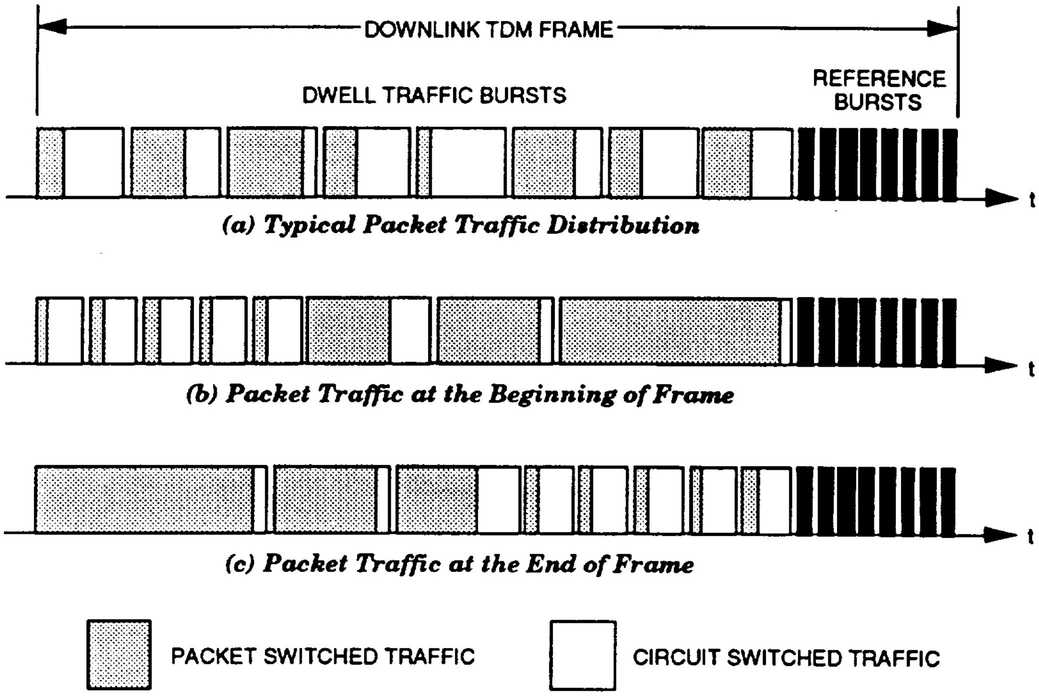
Figure 40. Shared Buffer for Circuit and Packet Switched Traffic

Between the two approaches for output buffering, the shared memory approach provides more flexibility and better memory utilization. For example, the given storage capacity can be dynamically allocated to circuit and packet switched traffic according to their traffic intensities. In an extreme case, the entire memory can be dedicated to one type of traffic, provided that the beam does not have the other type of traffic. The following discussions assume the shared memory approach.

4.2.2 Buffer Size

Circuit switched traffic generally requires at least one frame of buffering for rate conversion, time slot interchange (from uplink to downlink), and TDM formatting. Although a shorter TDM frame period is desirable in terms of hardware complexity, it decreases downlink frame efficiency. According to Table 3, the selection of a frame period of 0.5 ms results in a frame inefficiency of 4.17 percent and is considered for a baseline design. The buffer size required for supporting circuit switched traffic is a modest 5.12 kbytes per beam. One TDM frame corresponds to 32 bits for a 64-kbit/s traffic channel and 1,024 bits for a 2.048-Mbit/s uplink carrier.

The destination dwell areas for packet switched traffic are non-deterministic and randomly change from one frame to another. Thus, the buffer requirement for packet switched traffic must consider the impact of statistical distribution of packet destinations. To achieve a packet loss ratio of 10^{-9} at a traffic loading factor of 0.9 and uniform distribution, the buffer must accommodate about 96 packets per beam [18] which corresponds to 6.15 kbytes of storage for a packet size of 512 bits. Another factor to be considered is a staggered TDM burst operation to different dwell areas, as shown in Figure 41. Packet switched traffic may be concentrated at the beginning or the end of a downlink TDM frame. This implies that packets may be queued on the satellite for up



(a) Typical Packet Traffic Distribution

(b) Packet Traffic at the Beginning of Frame

(c) Packet Traffic at the End of Frame

Figure 41. Distribution of Packet Switched Traffic in Downlink TDM Frame

to 278 µs (81.92/147.5 x 0.5 ms) prior to downbeam transmission, and the buffer size for this queueing is 2.28 kbytes. Thus, the buffer requirement for packet switched traffic is 8.43 kbytes.

The total buffer size required for shared memory operation is the sum of the buffer sizes for circuit and packet switched traffic and is 13.6 kbytes. A memory configuration depends on the switch structure, switch speed, and memory access speed selected. A multistage banyan-based switch may require a switch speed of around 200 Mbit/s, resulting in a memory configuration of 16K x 8 or 8K x 16 with an access speed of 20 ns or 40 ns, respectively. For a high-speed optic ring, the same memory configuration requires an access speed of 5.5 ns (16K x 8) or 11 ns (8K x 16). The memory access speed can be reduced with a wider memory width or the use of ping-pong memories.

To ensure no data loss operation for circuit switched traffic, a storage space of 5.12 kbytes may be reserved. This space may also be used by packet switched traffic on a contingency basis. A temporarily leased space for packet switched traffic must be able to be vacated in about 270 ms for new circuit switched calls, if needed.

Multicast traffic channels (both circuit or packet traffic) may be replicated and stored in multiple memory locations along with their destination dwell area information. Alternately, replication may be performed at the time of downlink transmission. The former requires a larger memory space than the latter. However, there will be no significant impact on overall performance, since the allocated buffer size is large enough to handle the peak traffic volume (i.e., ~150 Mbit/s). In this regard, either replication method is acceptable.

The TDM controller monitors a packet queue for each dwell area and controls the amount of packet transmission within the allocated burst lengths. Because of the statistical nature of packet switched traffic, the distribution of a queue for different dwell areas is often uneven, and from time to time a queue length for some dwell area becomes significantly larger than for others. In this situation, the burst lengths to those areas with larger queues may be expanded by sending new time plans to the affected areas. The new downlink time plans will be implemented by the earth stations at a designated frame number. This procedure does not involve transmit traffic reconfiguration and can easily be implemented by the ANC within a few frame periods upon detection of a potential congestion state. Figure 42 illustrates a time plan switchover process. The dynamic capacity allocation based on dwell area queue status will relax an on-board congestion problem. A detailed analysis is recommended to quantify the improvement.

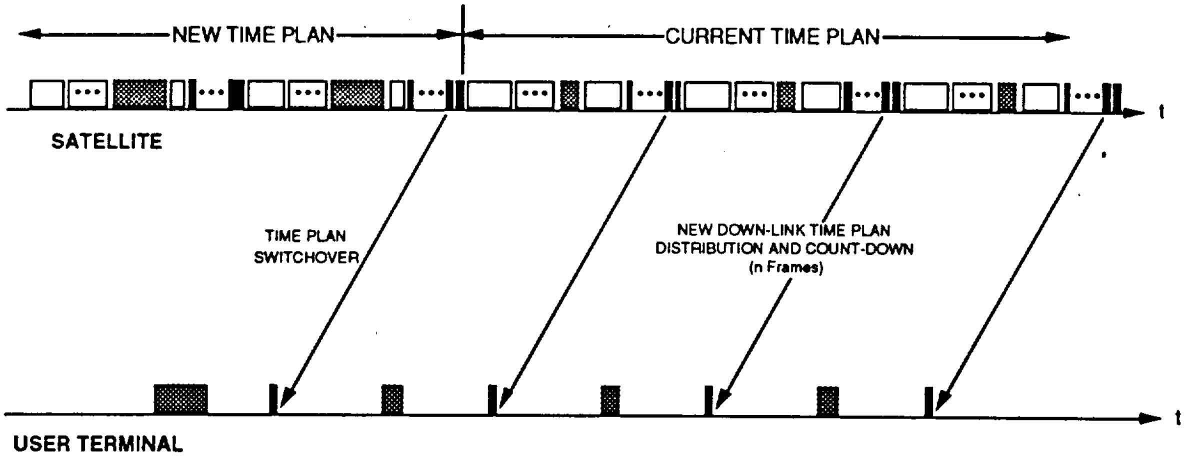


Figure 42. Downlink Time Plan Switchover Process

5 Integrated Circuit and Packet Switched System

On-board switching provides multimedia (voice, video, and data), multipoint (point-to-point, point-to-multipoint, and broadcast), and multirate services. In this section, an integrated switch is considered to provide unified switching/routing for both circuit and packet switched traffic. Compared with the two switching system scenario, the integrated switch has the following advantages. Integration simplifies the network management functions and makes the introduction of new services with different characteristics easier. It also provides simpler implementation and control, less hardware, easy fault tolerance and redundancy structures, reduced mass and power, and unified routing procedure. Most importantly, the integrated switch is more flexible in allocating the capacity of the switch between circuit and packet switched traffic. The following presents a design approach to an integrated switch using a multistage network.

The circuit traffic is segmented into packet formats at the sending stations and reassembled into channel formats at the receiving stations. Both packet and circuit data have the same packet format. The uplink uses the slotted transmission format. The unified packet format occupies one slot of uplink frame. The integrated access scheme uses the combination of TDM and packet transmission. (TDM is conventionally used for circuit switching while packet transmission uses statistically multiplexing.) An integrated switch can provide services for both circuit switched and packet switched data and, at the same time, preserve the QOS for each class. A space switch without the time stage is sufficient to route both circuit switched and packet switched traffic.

The packet structure contains the synchronization header, indication field, destination address, source address, control field, information payload, and forward error control (FEC). Note that the FEC can be used for synchronizing the packet as the cell delineation algorithm performed in ATM cell synchronization. In this case, the synchronization header is not required. Whether the routing tag of the packet should be prepended at the earth station or on-board the satellite largely depends on the link efficiency (routing tag size) and on-board translation table complexity. For point-to-point connections, since the size of the routing tag is very small (only Log N), it is suitable to generate the routing tags at the earth stations. For point-to-multipoint connections, the routing tag is at least N bit long. A trade-off has to be made to determine where to prepend the routing tag.

Since a fast packet switch routes a packet to the destined output port based on the routing tag, whether the packet contains circuit data or packet data will be transparent to the switch. Hence, the operation of the fast packet switch described in Section 3 can also be applied to the integrated switch. However, since circuit switched data comes into the system almost periodically and has a more stringent performance requirement (such as delay jitter) than the packet data, the operation of the switch has to be modified accordingly. Potential modification areas include a packet transfer scheduling algorithm and buffer space allocation. For output contention control, priority control is required to guarantee a certain high-QOS circuit switched data to pass through the switch faster than the other services. Priority control of the fast packet switch is performed during the packet transfer scheduling phase. If two packets are competing for the same link (internal blocking) or competing for the same output port (output contention), the packet transfer scheduling algorithm guarantees that a high priority packet will win over the low priority packet. If two packets have the same priority, they will be competing on an FCFS basis. A possible implementation of priority control for a point-to-multipoint nonblocking switching fabric with input buffering and input ring reservation scheme is described as follows [19].

The format of the tokens is modified to accommodate priority control. There are N tokens for the N output ports and there are N priority subfields for N tokens (see Figure 43). Whenever an input port reserves the output port, the priority of the packet waiting in the queue is also inserted in subfield $P_{i,}$ where i is the position of token i. Following the same procedure in the input ring reservation scheme, the packet at the next input port checks the availability of the output port. If the output port has been reserved, then the input port checks the priority level associated with this token. If the priority level is lower then its own priority level, the input port overwrites the priority field. If this occurs, the input port whose priority subfield has been overwritten needs to be notified. The notification scheme is very simple. After the packet transfer scheduling has been finished for all the input ports, the token stream is sent back to the input ports for confirmation. Every input port checks the priority subfield associated with the token to see if the priority is still the same as its own priority. If they are the same, confirmation is achieved and the packet can be transmitted at the beginning of the next slot. If they are different, it means some other packet with a higher priority at another input port has overwritten the token and the result is the low-priority packet has to retry the reservation request at the next slot time. In summary, the scheme circulates the tokens through the input ports twice. Loop 1 is for the input ports to reserve the output ports and loop 2 is for the input ports to confirm that the reservation of the output ports has been successful. The same principle can also be applied to the switch with a blocking switching fabric, which will not be repeated here.

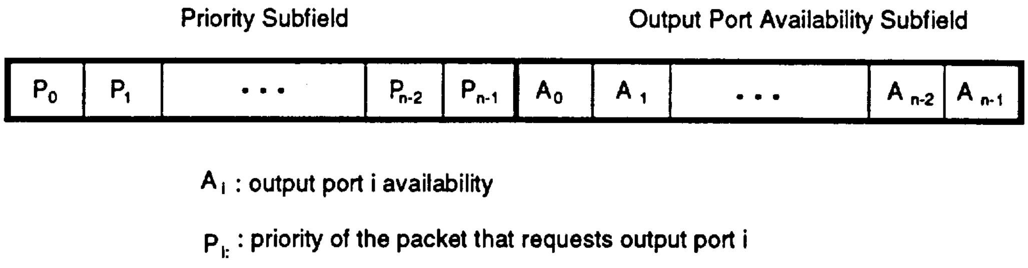

Priority Subfield
Output Port Availability Subfield
P_0
P_1
• • •
P_{n-2}
P_{n-1}
A_0
A_1
• • •
A_{n-2}
A_{n-1}
A_i : output port i availability
$P_{i:}$: priority of the packet that requests output port i

Figure 43. Token Format with Priority Subfield

Since circuit traffic comes into the system almost periodically, the buffer space for the circuit data can be reserved in advance. In essence, the buffer space has been divided into two portions. One portion is to reserve for the circuit switched data while the other portion is shared by the packet switched data. Congestion control is required only for packet switched data. The input buffer space for circuit switched data will be very small since the maximum delay encountered in the switch cannot exceed the QOS performance requirement. Most input buffer space is used for packet switched data. To have an effective priority control, the buffer space may be completely divided so that HOL blocking of one traffic type will not affect the other traffic type.

6 Approaches to Congestion Problems

Congestion occurs when the demand for resources in the network, in this case the on-board information switching processor (ISP), exceeds the capacity. Circuit-switched traffic alone does not induce congestion.because the capacity is scheduled ahead of the time the traffic arrives. With packetized traffic flowing through the switch, congestion is inevitable, and proper congestion control techniques must be employed.

There are a whole range of traffic management methods by which congestion in packet networks can be avoided and controlled. The network traffic needs to be characterized, and proper congestion control methods must be applied to the network according to the traffic characteristics.

The first traffic management method is the call or connection admission function. This is an integral part of dynamic resoure assignment. When a request for a new call or connection is received, this function decides to either accept or reject the request. If the decision is to accept the call, the network ensures the availability of adequate bandwidth based on the traffic characterization of the call (mean bit-rate, peak bit-rate, mean holding time, peak burst duration,etc.) and the quality of service (QOS) requirements (packet loss rate, maximum allowable delay, etc.). The acceptance decision also reflects the fact that the existing calls within the system will continue to meet the QOS requirements without degradation. The call rejection decision reflects the fact that such guarantees are not possible either for the new call or for the existing call or for both. For circuit-switched traffic and for constant bit-rate services, this is relatively easy. But for other types of traffic, this is difficult. It is obvious that a conservative call admission function will reduce the level of congestion (not completely eliminate it) but also lower the effective utilization of the bandwidth. The goal is to maintain a high level of utilization by accepting the maximum number of calls possible and by managing the resulting congestion. This function may be implemented either on-board the satellite with direct access to the buffer status of different downbeams, or at the central management control center, or at the entry points to the earth stations. Each location has its own advantages and disadvantages, and a tradeoff is needed.

Once calls have been admitted into the system, monitoring is necessary to ensure that the incoming traffic conforms to rates expected by the system, which may be the average bit-rate, peak bit-rate, peak burst duration, etc. There needs to be a traffic flow control or connection control function of the input traffic. This may be most easily done at the access points of the network. This may be achieved by various means, by outright discarding packets that exceed a certain threshold, by buffering and smoothing the traffic stream to the desired rate within acceptable delay and jitter, by introducing spacers with the leaky bucket algorithm. However, for finer control with greater bandwidth utilization, packets in violation can be tagged lower priority within certain limits and be allowed into the network with the assumption that in case congestion occurs on the path, packets in violation will be discarded first. The ISP must perform some form of traffic enforcement function. This needs an appropriate queueing scheme to ensure: (a) packets not in violation have greater priority than tagged packets and (b) if the network is not adversely affected and the QOS requirements of existing connections met, the tagged packets are not discarded and delivered.

In spite of good call admission control and call parameter control, it is possible for congestion to develop at intermediate points of a route. To combat this, so-called reactive control has to be employed. These are explicit congestion notification with source throttling based on either rate-based control or window-based control, in-call parameter renegotiation. There exist various leaky bucket algorithms that have been employed in the past with packet networks. Another method is to drop packets selectively at intermediate nodes. The selection can be based on tagging packets on violation as discussed before or by assigning a priority scheme.

Based on a number of techniques for congestion control, a thorough study of the entire system and the expected traffic is necessary to determine the functions to be performed on-board the satellite and at the ground stations. The specific implementation schemes for these functions could range from the various classical methods to certain recently developed neural network techniques.

7 Conclusion

The contention problem in destination-directed packet switching, as investigated in detail in this report, is not a major concern. It can be completely avoided using a contention-free switch architecture. The use of a fiber optic ring, for example, should be able to provide a contention-free switching function for a total system capacity of about 2 Gbit/s. As technology in optic devices and high-speed semiconductor devices progresses, the total switching capacity can be significantly increased.

Contention, however, is an inherent property of multi-stage switching networks. The techniques to resolve contention include output port reservation and path setups prior to packet routing. These techniques reduce a switch throughput by 20 to 40 percent due to scheduling efficiency and contention of path setup packets and require an increase in switching speed by 25 to 67 percent to maintain the desired switch throughput. This is based on statistically independent packet transmission from uplink beams to different downlink beams; otherwise, although this is very unlikely, the switching speed must be increased further. With an increased switching speed, the contention problem will virtually disappear.

Another technique to resolve contention employs dedicated paths from each input port to different output ports and address filtering. Contention occurs at the output concentrator because of a limited buffer size. Since a switch fabric is contention-free, this problem may be regarded as a congestion problem.

Congestion is a more difficult problem associated with destination directed packet switching. This problem is not unique to satellite communications and in fact has been extensively studied for terrestrial ATM networks. Some of the techniques proposed for terrestrial networks may not be effectively used for satellite networks because of long propagation delay. To alleviate the impact of this delay on the classical congestion control methods, the predictive techniques using neural network formulation may need to be employed. A congestion control procedure must be devised as a part of overall network control, including packet queue monitoring and buffer management by the on-board processor and user earth stations, call admission control at the user and network levels, and satellite capacity allocation procedures. It is recommended that the congestion problem be investigated in a future study, since circuit switching, as an alternative to destination directed packet switching, is simply inadequate for packet switched traffic.

Another result from this study indicates that a fast packet switch can support both circuit and packet switched traffic. Destination directed packet switching for circuit switched traffic requires additional processing, such as packet assembly, bit interleaving, and header error checking, but it eliminates control memories, memory update processing, switchover coordination, and a path finder procedure for channel routing. The benefits gained from this conversion can be substantial. A detailed study for an integrated network architecture for circuit and packet switched traffic is strongly recommended. The study should cover specific network requirements, frame and packet structures, frame efficiency, detailed baseband processor block diagram designs (including all the necessary functions from MCD output to modulator input), acquisition and synchronization, capacity request/allocation procedures, flow/congestion control procedures, and earth station block diagram designs. The key to the effective study will be well defined network requirements in the earlier phase of the study task.

8 References

[1] On-Board Processing Satellite Network Architecture and Control Study, Final Report, NASA Contract NAS3-24886, Prepared by COMSAT Laboratories, June 1987.

[2] S. J. Campanella, B. A. Pontano, and H. Chalmers, "Future Switching Satellite," AIAA 12th International Communication Satellite Systems Conference, Virginia, pp. 264-273, March 13-17, 1988.

[3] W. D. Ivancic and M. J. Shalkhauser, "Destination Directed Packet Switch Architecture for a 30/20 GHz FDMA/TDM Geostationary Communication Satellite Network," Second NASA Space Communications Technology Conference, Cleveland, Ohio, November 12-14, 1991.

[4] T. Inukai, D. J. Shyy, and F. Faris, "On-Board Processing Architectures for Satellite B-ISDN Services," Second NASA Space Communications Technology Conference, Cleveland, Ohio, November 12-14, 1991.

[5] M. Karol, M. Hluchyj, and S. Morgan, "Input vs Output Queueing on a Space-Division Packet Switch," IEEE Trans. on Communications, vol. 35, pp. 1347-1356, Dec. 1987.

[6] G. Dorazza and C. Raffaelli, "Acknowledegement-Based Broadband Switching Architectures," Electronics Letters, vol. 25, no.5, pp.332-334, 1989.

[7] L. R. Goke and G. J. Lipovski, "Banyan Networks for Partitioning Multiprocessing Systems," First Annual Computer Architecture, pp. 21-28, 1973.

[8] K. E. Batcher, "Sorting Networks and Their Applications," AFIPS , vol.32, pp. 307-314, 1968.

[9] B. Bingham and H. Bussey, "Reservation-Based Contention Resolution Mechanism for Batcher-Banyan Packet Switches," Electronic Letters, vol. 24, no. 13, pp. 772-773, June 1988.

[10] K. W. Sarkies, "The Bypass Queue in Fast Packet Switching," IEEE Trans. on Communications, vol. 39, no. 5, pp. 766-774, May 1991.

[11] N. Arakawa, Akira Noiri and H. Inoue, "ATM Switch for Multi-Media Switching System," ISS, vol. 5, pp. 9-14, 1990.

[12] A. Cisneros, "Large Packet Switch and Contention Resolution Device," ISS, vol. 3, pp. 77-83, 1990.

[13] A. Hunag and S. Knauer, "Starlite: A Wideband Digital Switch", IEEE GLOBECOM, pp. 121-125, 1984.

[14] J. Y. Hui and E. Arthurs, "Broadband Packet Switch for Integrated Transport", IEEE JSAC, vol. 5, no. 8, pp. 1264-1273., Oct. 1987.

[15] D.-J. Shyy, "Nonblocking Multicast Fast Packet/Circuit Switching Networks," COMSAT Invention Disclosure No. 31-E-10, June 1991.

[16] K. Y. Eng, M. G. Hluchyj and Y. S. Yeh, "Multicast and Broadcast Services in a Knockout Packet Switch," IEEE INFOCOM, pp. 29-34, 1988.

[17] G. Nathan, P. Holdaway, and G. Anido, "A Multipath Multicast Switch Architecture," 1988.

[18] Y. Shobatake, et. al., "A One-Chip Scalable 8 x 8 ATM Switch LSI Employing Shared Buffer Architecture," IEEE Journal on Selected Areas in Communications, vol. 9, no. 8, October 1991, pp. 1248-1254.

[19] T. Lee, M. Goodman, and E. Arthurs, "A Broadband Optical Multicast Switch," ISS, vol. 3, pp. 7-13, 1990.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE July 1995	3. REPORT TYPE AND DATES COVERED Final Contractor Report

4. TITLE AND SUBTITLE

Information Switching Processor (ISP) Contention Analysis and Control

5. FUNDING NUMBERS

WU-506-72-21
C-NAS3-25933

6. AUTHOR(S)

Thomas Inukai

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

COMSAT Laboratories
22300 Comsat Drive
Clarksburg, Maryland 20871

8. PERFORMING ORGANIZATION REPORT NUMBER

E-9657

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135-3191

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA CR-195471

11. SUPPLEMENTARY NOTES

Project manager, Heechul Kim, Space Electronics Division, NASA Lewis Research Center, organization code 5650, (216) 433-8698.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Category 17

This publication is available from the NASA Center for Aerospace Information, (301) 621-0390.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

In designing a satellite system with on-board processing, the selection of a switching architecture is often critical. The on-board switching function can be implemented by circuit switching or packet switching. Destination-directed packet switching has several attractive features, such as self-routing without on-board switch reconfiguration, no switch control memory requirement, efficient bandwidth utilization for packet switched traffic, and accommodation of circuit switched traffic. Destination-directed packet switching, however, has two potential concerns: (a) contention and (b) congestion. And this report specifically deals with the first problem. It includes a description and analysis of various self-routing switch structures, the nature of contention problems, and contention and resolution techniques.

14. SUBJECT TERMS

Information switching processor; Contention analysis

15. NUMBER OF PAGES

86

16. PRICE CODE

A05

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT